

AUTOMATIC FOOTBALL VIDEO HIGHLIGHTS EXTRACTION

Jan Erik Voldhaug Stian Johansen Andrew Perkis

Centre for Quantifiable Quality of Service in Communication Systems*
Norwegian University of Science and Technology
Trondheim, Norway
{voldhaug,stianjo,andrew}@q2s.ntnu.no

ABSTRACT

This paper defines a system for automatic analysis of, and highlights extraction from football television footage. The analysis is based solely on visual or cinematic information in the video, using shot type classification and colour analysis on video frames. Two different versions of the system have been developed, enabling work on decoded video frames as well as direct analysis of the compressed bitstream. The latter scheme processes the DC image of the I-frames, rather than pixels in decoded frames. Testing shows promising results, with shot type classification accuracy of just over 90 % and excellent run time.

1. INTRODUCTION

In today's society time is, for many, a limited resource. With ever more people getting connected to the Internet and with high bandwidth connections becoming more and more widespread, the amount of information available has never been larger. The last years have seen a substantial increase in hardware performance and decrease in storage media cost. Together with the development of multimedia compression standards, this contributes to widespread exchange of multimedia content. High speed Internet connections also enable applications such as broadcasting over IP. With time being limited, there is a limit as to the amount of information each person can process. Information to be processed has to be selected.

Sports video appeals to large audiences. In Europe and South America particularly football, also known as soccer, is extremely popular. Football (and sports) videos are special in the way that important events in a game in general only occupy a small portion of the total content. Hence they can be significantly shortened without losing the main information content.

In this paper we define a system for analysis and summarization of football footage. The system uses shot classification based on colour analysis. The same scheme is used in [2] to identify breaks in games. [1] proposes a framework for football video summarization where colour analysis and shot classification is combined with object based features such as goal detection, referee detection and penalty-box detection to improve performance. [4] uses recognition of playfield zones and camera motion analysis. [3] tracks movement of the ball and identifies game

features such as players and goal posts. [7] uses audio features to identify highlights in football as well as baseball and golf.

The work presented here is based on the observation that football playing fields are green. Two different schemes are used to access colour information from the frames. Whereas in the first approach pixels are accessed from decoded frames, the second approach does not include full decoding of the video. Instead, DC values of I-frames are accessed directly from the compressed MPEG-1 bitstream. The DC value is the zero-frequency component obtained from the discrete cosine transform, DCT, of MPEG-1. It represents the average value of a block. The DC values of all blocks in a frame comprise the DC image of the frame.

DC images of MPEG-1 I-frames are also the basis for colour analysis in [2]. In [5] the colour histogram of the DC images is used to detect scene changes. [6] extracts the MPEG *MBType* variables which hold information about the coding type of each macroblock. The amount of inter-frame prediction in each frame indicates the probability of a scene change.

Although this system primarily has been developed to summarize football video on the user end (on a set-top box, for example) similar processing and summarization can also be used to deliver (sports) video over narrow band networks. Adaptation of multimedia content to allow delivery to different users with different terminal characteristics and communication infrastructure is the focus of UMA [8].

This paper is organized as follows: An overview of the proposed highlights extraction system is given in sections 2 and 3. Results from testing are given in section 4.

2. HIGHLIGHTS EXTRACTION FROM DECODED VIDEO

This section deals with analysis on pixels in decoded frames. The main structure of the system is shown in Figure 1.

For this work, the standard RGB colour representation is not convenient. Instead the RGB values of decoded frames are transformed to corresponding coefficients in the HSB colour space, before analysis. HSB separates *hue*, *saturation* and *brightness* into three different parameters. *Hue* determines the dominant wavelength of the colour with values ranging from 0 to 360 degrees. *Brightness* describes the level of white light (0 - 100 %), whilst *Saturation* describes the proportion of chromatic element in a colour. Values range from 0 to 100 %, where low values indicate that the colour has much "grayness" and

*"Centre for Quantifiable Quality of Service in Communication Systems, Centre of Excellence" appointed by The Research Council of Norway, funded by The Research Council, NTNU and UNINETT. <http://www.q2s.ntnu.no/>

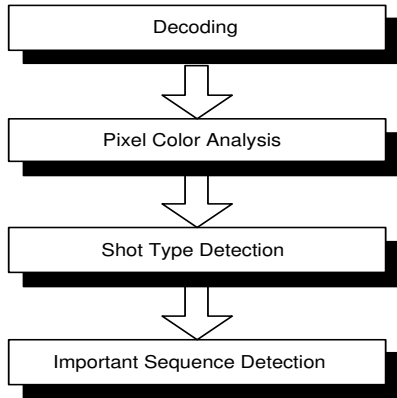


Figure 1: Key system elements - Analysis on decoded video.

will appear faded. As humans are much more sensitive to hue than to saturation and brightness, one parameter becomes far more important than the others and the HSB representation is therefore excellent for colour analysis.

As there is a strong correlation between consecutive frames, all frames need not be analyzed. For frame rates of 25-30 frames per second, which are typical, we found that it is sufficient to analyze every fifth frame. Information about faster variations is not vital to the algorithm. Nor does the shot type change several times per second.

2.1. Pixel Colour Analysis

The colour analysis is in many respects the foundation of the entire system. Depending on certain minimum and maximum limits for the HSB values, each pixel is labeled either *green* or *not green*. The limits used are derived from colour theory, and adjusted after testing. As discussed earlier, the hue parameter is by far the most important of the three in colour detection and we assume that the football playing field is green. From theory it is known that a hue value of $1/3$ (or 120 degrees) equals the colour green. The exact colour of the playing field will obviously differ between different stadiums, but there will also be variations within one particular playing field. These variations are caused by differences and changes in light (weather, floodlights etc.), different grass quality and possibly grass types in various parts of the pitch. Other parameters that will influence colours include the camera (lenses / optics), A/D conversion and compression scheme. Therefore the acceptance limits must be set wide enough to allow for the different shades of green.

Minimum and maximum limits for hue have been set at $1/3 - 0.15$ and $1/3 + 0.15$, respectively (± 54 degrees). For the other two parameters, saturation and brightness, minimum limits are set to 0.45 and 0.18 respectively. These values have emerged from testing, and the purpose of these limits is to ensure that too dark colours are not recognized as green. If the saturation and / or brightness values become too low, the hue parameter becomes irrelevant and the colour can no longer be determined.



Figure 2: Shot type examples. From left to right; long shot, medium shot and out of field shot.

2.2. Shot Type Classification

The shots used in a football production can be classified into three classes [1], [2]. These are *Long shots*, (*In-field*) *Medium shots* and *Close-ups / Out of field shots*. A long shot displays a global view of the field, shot from a camera high in the stands. This is the most frequently used shot type. Medium shots are zoomed in shots showing most of one or more bodies. A close-up shot is a close view (above waist) of one player or official. An out of field shot shows something not on the playing field, e.g. coaches, substitute players or spectators. Close-ups and out of field shots both indicate a break in the game.

For each frame to be analyzed the number of green pixels are counted, and based on the number of green pixels present the shot type is determined. To be able to separate the three shot types, two limits are needed. These limits refer to the ratio between green pixels and the total number of analyzed pixels in a frame, and have been set to 0.20 (medium-limit) and 0.45 (long-limit) in this work. The values are chosen based on testing and values found in other studies. [1] suggests ratios of 0.10 (10 % green pixels) and 0.40 (40 % green pixels) as initial minimum limits for medium and long shots respectively. Similarly [2] proposes 0.10 and 0.50. Comparing the amount of green pixels in each frame to these limits thus decides the shot type. If the amount of green pixels is larger than the long-limit, the frame is labeled as a long shot. If the amount of green pixels is not larger than the long-limit but larger than the medium-limit, the frame is labeled as a medium shot, and so on.

To improve performance further, the system weights different pixels differently depending on their position within the frame. This weighting is done by partitioning each frame horizontally and vertically in 3:5:3 proportion into 9 non-overlapping rectangular sections. A pixel is treated according to which section it is in.

Examples of errors that can be avoided using the weighting scheme includes long shots of action on the far side of the playing field from the camera position. A long shot of this will produce a frame with less green pixels as much of the top area will show the stand and spectators instead of grass. Hence such shots can very easily be misinterpreted as medium shots. Another very common error is to mistake a medium shot, for instance of one body with grass around, for a long shot.

2.3. Important Sequence Detection

The algorithm is based on the observation that when an important event occurs in a football match, the television production changes its use of camera angles for a period of time. The long shots, which are predominant when the ball is in normal open play, are replaced by medium shots,

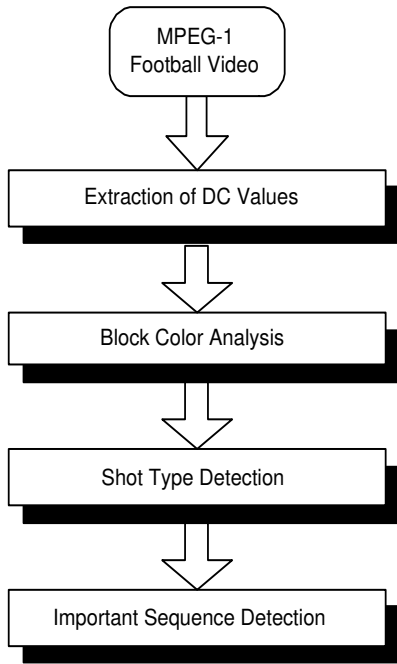


Figure 3: Key system elements - Analysis on compressed video.

close-ups of players, referees and / or coaches as well as out of field shots. This assumption holds for multi-camera productions and can be easily observed. Thus, the search for an important sequence equals the search for a series of consecutive or near consecutive non-long shot frames. If such a series longer than a given limit is found, then this sequence of frames will be included in the highlights.

3. ANALYSIS ON COMPRESSED VIDEO

Computational cost and run time are important parameters, and it is therefore desirable to eliminate the need for full decoding. In this approach we perform shot type classification and hence important sequence detection directly on the compressed bitstream.

The fact that we can eliminate the decoding from our system significantly reduces the complexity and running time.

3.1. Extraction of DC Values

This approach is based on MPEG (-1) compressed video [9]. In MPEG-1 video, each frame is divided into 16×16 pixel macroblocks with each macroblock containing 6 8×8 blocks; 4 luminance blocks and 2 chrominance blocks. These blocks are DCT transformed individually. The DCT results in 64 coefficients; 63 AC values and 1 DC value. The DC value is the zero-frequency component. It is proportional to the average value of its block. Together the DC values of all the blocks in a frame form a DC image. Figure 4 shows an example of a DC image.

Start codes and header information from the different layers in the bitstream are used to locate the blocks and their coefficients. All DC values from the frames are collected and combined to form the DC image.



Figure 4: DC image example. Each 8×8 pixel block has one DC value, hence the dimension of the DC image of a 352×288 pixel image is only 44×36 . Enlarged version included for illustration purposes.

In this work only extraction of DC values from I-frames has been implemented. The reason for this is primarily computational complexity. P- and B-frame DC values can also be extracted in the same manner, except temporal coding has to be taken into consideration. Obviously, simply ignoring B and P frames is not by any means ideal but as the I-frame rate typically is about one in every 10-15 frames or about 2 per second (25 fps), this should be sufficient.

3.2. Block Colour Analysis

The use of the YUV colour space enables subsampling of chrominance. In MPEG-1 the chrominance components are subsampled by a factor of 4, which means that there is only one Cb (U) and one Cr (V) component for every 4 luminance (Y) components. To compute the colour of a block, the luminance value of the desired block is combined with the DC values of the chrominance blocks in the same macroblock. The resulting YUV triplet is converted to RGB and finally to the HSB colour space. As described in chapter 2 the HSB value of every block is checked and depending on decision limits the blocks are labeled either green or not green. The concept and the limits used are the same as in chapter 2. The only difference is that instead of working on pixels, the analysis is now performed on blocks.

Shot type classification and important sequence detection is also done in the same way as described in chapter 2.

4. RESULTS

To evaluate the performance of the system, two different implementations were created and run on an actual football video. The test video is in MPEG-1 format with a frame rate of 25 frames per second and 352×288 pixel resolution.

Analysis results for parts of the video were extracted and manually compared to the actual frames. For both implementations (with and without decoding) the amount of analyzed frames equals about 15 minutes of video in the case without decoding, (only I-frames are processed), and about half of that in the case with full decoding (every 5th frame is processed). Hence the actual frames that are used in this test are not all the exact same for both cases. This is not ideal, but even so the number of frames should be

	Long	Medium	Close-Up
Long shots	95 %	5 %	0
Medium shots	16 %	82 %	2 %
Close-ups/ Out-of-field	2 %	13 %	85 %

Table 1: Shot type classification - full decoding.

	Long	Medium	Close-Up
Long shots	95 %	5 %	0
Medium shots	18 %	76 %	6 %
Close-ups/ Out-of-field	0 %	3 %	97 %

Table 2: Shot type classification - analysis on DC values.

sufficient to give an indication of the performance. Table 1 and 2 show the occurrence of false and correct recognitions for each shot type in the test sequence and what types the system recognized them as. The tables should be read horizontally. Each row shows how many of a given shot type that is recognized as a long shot, a medium shot and a close-up / out-of-field shot respectively.

The total number of errors for each implementation is shown in table 3. The value *adjusted accuracy* in the bottom row of the table emerges when errors that don't have any influence on the important sequence detection are ignored.

The important sequence detection is harder to evaluate, as defining what an important sequence is, to an extent, is subjective. Both implementations successfully detected the two goals in the video. However, the implementation without full decoding missed a couple of key situations and also had more dubious / false inclusions.

In terms of shot type classification the system seems to perform very well, with the full decoding implementation performing slightly better than the other. Accuracies of 91.42 % and 91.24 % with and without full decoding, respectively, are very encouraging. When ignoring errors that are irrelevant to the system, accuracies increase to 93.22 % and 92.81 %. Ekin et. al. [1] report view classification results just under 90 %. It must, however, be stressed that testing on this system has so far been limited.

5. CONCLUSION

This paper has presented a system for automatic football video highlights extraction, and shown that this can be done with simple colour analysis both on decoded and compressed video. Testing shows promising results, with over 90 % accuracy in shot type classification. The implementation using full decoding offer slightly better performance in terms of important sequence identification, with both versions performing reasonably well. When implemented without full decoding, the system has low complexity and excellent run time, capable of running at only a fraction of real-time.

Method	Analysis on Decoded video	Analysis on Compressed video
Shots	2169	2169
Correct	1983	1979
False	186	190
Accuracy	91.42%	91.24%
Significant errors	147	156
Adjusted accuracy	93.22%	92.81%

Table 3: Shot type classification accuracy. Correctly and erroneously recognized shots for both implementations.

6. REFERENCES

- [1] A. Ekin, A.M. Tekalp, R. Mehrotra, *Automatic Soccer Video Analysis and Summarization*, Image Processing, IEEE Transactions on , Volume: 12 , Issue: 7 , July 2003
- [2] P. Xu, L. Xie, S-F. Chang, A. Divakaran, A. Vetro, H. Sun, *Algorithms and System for Segmentation and Structure Analysis in Soccer Video*, Multimedia and Expo, 2001. ICME 2001. IEEE International Conference on , 22-25 Aug. 2001
- [3] D. Yow, B-L. Yeo, M. Yeung, B. Liu, *Analysis and Presentation of Soccer Highlights from Digital Video*, Proceedings Asian Conference on Computer Vision (ACCV), 1995
- [4] J. AssfaIg, M. Bertini, C. Colombo, A. Del Bimbo, W. Nunziati, *Automatic extraction and annotation of soccer video highlights*, Image Processing, 2003. Proceedings. 2003 International Conference on, Volume: 2, Sept. 14-17, 2003
- [5] K. Shen, E. J. Delp, *A Fast Algorithm for Video Parsing Using MPEG Compressed Sequences*, Image Processing, 1995. Proceedings., International Conference on, Volume: 2 , 23-26 October 1995
- [6] Calic, J., Izquierdo, E., *Efficient Key-Frame Extraction and Video Analysis*, Information Technology: Coding and Computing, 2002. Proceedings. International Conference on , April 8-10, 2002
- [7] Xiong, Z., Radharkrishnan, R., Divakaran, A., Huang, T. S., *Audio Events Detection Based Highlights Extraction from Baseball, Golf and Soccer Games in a Unified Framework*, Multimedia and Expo, 2003. ICME '03. Proceedings. 2003 International Conference on , Volume: 3 , July 2003
- [8] A. Perkis, Y. Abdeljaoued, C. Christopoulos, T. Ebrahimi, J. F. Chicharo, *Universal Multimedia Access from Wired and Wireless Systems*, Circuits Systems Signal Processing, Vol. 20, No. 3, 2001
- [9] ISO/IEC JTC1/SC29/WG11 International Standard 11172; "Coding of moving pictures and associated audio for digital storage media up to about 1,5 Mb/s", November 1993