AUTOMATIC KEY-FRAME EXTRACTION FROM BROADCAST SOCCER VIDEOS

Nielsen C. Simões[1], Neucimar J. Leite[2] and Beatriz Marcotegui[3]
[1]*University of Mato Grosso do Sul (UEMS), Dourados, MS, Brazil*

[2]*Institute of Computing, University of Campinas (UNICAMP), Campinas, SP, Brazil*

[3]*Centre de Morphologie Mathématique (CMM), Ecole des Mines de Paris, Fontainebleau, France*

Keywords: Key-frame, Video analysis, Shot classification, Broadcast soccer video, Visual rhythm.

Abstract: This paper presents a new approach for broadcast soccer video navigation and summarization based on specific representative images of the video. It also takes into account some soccer video features to better describe these videos. This work considers a special color reduction based on an HSV subquantization and a shot classification approach for soccer videos by exploring the dominant color related to the playground area.

1 INTRODUCTION

The manual analysis and annotation of video databases is a long and arduous task which can generate mistakes due to the operator weariness, for example. Automatic analysis is an important task for video semantic understanding. A digital video segment is defined as a sequence of images or frames. A shot is an uninterrupted subset of frames recorded from the same camera. Shot detection is one of the first tasks in automatic analysis of videos. Video edition generates some transitions between shots. There are many approaches for shot transition detection concerned with different kinds of videos, such as commercials, news and movies.

Key-frames are usually defined to represent a shot and are commonly used as video edition tools which deal with shots in the time-line video segment. Some video indexing techniques and analysis also consider key-frame information in order to reduce the amount of data to be analyzed.

A key-frame is a representative image of a shot and its extraction is an important tool for the process of video semantic analysis (Ciocca and Schettini, 2005; Doulamis et al., 2000; Dufaux, 2000; Wolf, 1996). In general, it is frequently used for video visualization (Arman et al., 1994; Komlodi and Marchionini, 1998; Tse et al., 1998; Zhong et al., 1996) and recovery (Arman et al., 1994; Liu et al., 2003; Sze et al., 2005; Pardo, 2006), or even as a tool for scenes detection (Xiong et al., 1997).

Most key-frame detection approaches consider, as representative images, the frames obtained from a predefined position in each shot (Arman et al., 1994; Ueda et al., 1991; Zhang et al., 1995). This simple method can be efficient for short videos such as TV advertisements. Other approaches are based on dissimilarity measures between consecutive frames (Ciocca and Schettini, 2005; Doulamis et al., 2000; Yeung and Liu, 1995; Xiong et al., 1997) or also by computing the local minimum related to image motion (Dufaux, 2000; Wolf, 1996).

This work presents a new approach for automatic key-frame extraction from TV broadcast soccer videos, useful for video browsing and navigation. It is also proposed an specific video representative image, a special color reduction and a shot classification for soccer videos by exploring the playground color frequency. Section 2 presents a new approach for key-frame extraction. Section 3 shows experimental results using some Brazilian TV broadcast soccer videos. Section 4 draws some conclusions and future works.

2 KEY-FRAME EXTRACTION

(Sze et al., 2005) take into account the temporal histogram for each pixel in a group of frames related to

C. Simões N., J. Leite N. and Marcotegui B. (2009).
AUTOMATIC KEY-FRAME EXTRACTION FROM BROADCAST SOCCER VIDEOS.
In *Proceedings of the Fourth International Conference on Computer Vision Theory and Applications*, pages 216-223
DOI: 10.5220/0001805702160223

a given shot. In such case, the obtained key-frame is a composition of different frames which yields a key-frame representation that can be useful for video retrieval but not for video browsing and navigation, since this synthetic key-frame is not necessarily included in the set of the original frames belonging to the corresponding shot. (Pardo, 2006) also considers a video segment or a group of frames to compute the pixel-wise histogram which, again, can be useful for video retrieval but not for video browsing and navigation.

A simple key-frame extraction can be obtained by defining the first frame of each shot as representative image. Since soccer videos have a lot of long shots, some of these unique key-frames may not represent them properly. Thus, it might be desirable to select more than one key-frame associated with the content of these long duration shots. In general, a shot detection step must be considered before performing the key-frame extraction method. There are different shot detection approaches in the literature (Brunelli et al., 1999; Guimarães et al., 2003; Kim et al., 2001; Koprinska and Carrato, 2001; Ngo et al., 1998; Simões, 2004; Zhang et al., 1993), and new ones have already been introduced, mainly due to the specific characteristics of the various types of existent videos.

This work focuses on the key-frame extraction of soccer video images. Since the shot detection is beyond the scope of this paper it assumes the availability of an approach for shot detection, such as pixel-wise comparison (Brunelli et al., 1999; Koprinska and Carrato, 2001; Patel and Sethi, 1996; Zhang et al., 1993), the one considered for this task in this work. The key-frame extraction proposed here attempts to define at least one key-frame for each single shot. In cases of long duration shots more than one key-frame is defined to better describe these specific shots.

In TV broadcast soccer videos most shots are represented by large playground regions. In this sense, we propose a special color reduction approach, as well as a specific visual rhythm transformation and a method for shot classification by exploring this playground color information of the shots. As we will show in Section 2.3, this information will be used to identify shots which are closely related to the field, i.e., frames whose pixels yield high playground color density. The next subsections discuss in details each step of this key-frame extraction approach.

2.1 Color Reduction

A soccer match is played on a playground which, for processing purposes, can be considered as the background of the video images. In TV broadcast videos, in which a match is recorded through different cameras and angles, the playground information is presented most of the time, making the color information the representative or dominant feature in the corresponding shots. In order to estimate this color representativeness of the frames, a simple histogram operation can be used. Since this step consumes much memory and is sensitive to brightness conditions, for example, a special color reduction model is proposed to avoid these problems and also discriminate the playground color information.

By considering the HSV color space, the method performs a subsampling of its main components. Informally, the proposed approach reduces the *Hue* component to six values (primary and secondary colors, i.e., red, green, blue, yellow, cyan and magenta) while the *Saturation* component is reduced to eight values. This same reduction procedure is applied to the *Value* component (see Figure 1). All colors with low *Saturation* are reduced to eight gray levels, and all the other colors with low *Values* are represented by black. For example, due to changes in the time conditions the playground color can present significant brightness variations. Moreover, the playground grass can be of different types, yielding changes to the playground color saturation, as well. Thus, we can not properly consider brightness and saturation values for the playground color identification.

Here, the first, third and fifth *Hue* values of the color space are related to the primary colors (red, green and blue) and the second, fourth and sixth colors are related to the secondary colors (yellow, cyan and magenta). After the analysis of matches with different time and weather conditions the second *Hue* value, closely related to the actual colors of the playground, is associated to the playground color. *Saturation* and *value* components was kept fixed in the reduction approach for this hue information, which results in only one fixed color for the playground information. The other five *Hue* values can be reduced by considering seven different values for the *Saturation* component (low saturations are represented as gray levels) and seven different values for the *Value* component (low values are represented as black color). This maps all HSV values to one playground color, one black color, seven gray levels and 245 different colors: 5 (the 6 original *Hue* components minus 1 of the fixed *Hue* component) x 7 (values for *Saturation*) x 7 (values for the *Value* component). As a result, the method makes a reduction of a 24-bit color representation to 254 colors with only one of them related to the playground information.

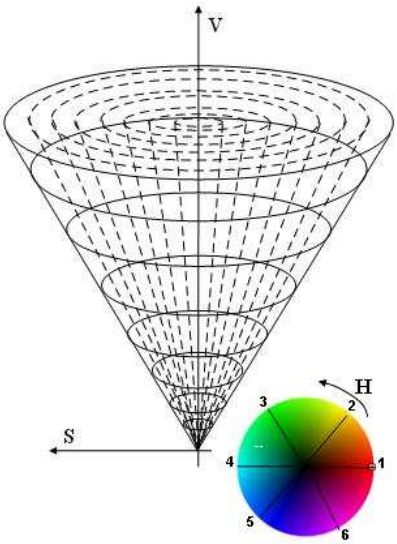

Figure 1: HSV partition for color reduction.

2.2 Visual Rhythm Approach

Visual Rhythm is a representative image of a whole video proposed firstly as a tool for shot detection approaches (Chung et al., 1999; Bezerra and Leite, 2003; Guimarães et al., 2003; Kim et al., 2001). The original visual rhythm (Definition 2.1) uses a transformation function from the spatial domain of the sequence ($2\mathbb{D}+t$) to $\mathbb{D}+t$.

Definition 2.1 Visual Rhythm. *Let $f_t(x,y)$ be the color value of the pixel (x,y) of a frame in time t, from a digital video with N frames. Let H and W be, respectively, the height and width of the frames. The visual rhythm image I_{VR} is the result of the following transformation:*

$$I_{VR}(t,z) = f_t(r_x \times z + a, r_y \times z + b),$$

where $z \in [0, H_{VR}-1]$ and $t \in [0, N-1]$, H_{VR} and N are, respectively, the height and width of the visual rhythm image; r_x and r_y represent a pixel subsampling rate, while a e b, a translation into each frame.

Definition 2.1 corresponds to a subsampling of the video, resulting in a set of representative image which keeps some of its main temporal features. A more general definition of this visual rhythm is given in Definition 2.2, in which a transformation function can be defined considering both local and global informations of the image sequences. Examples of typical local information for the visual rhythm representation are the center vertical or horizontal lines, or the main diagonal of the video frames.

Definition 2.2 General Visual Rhythm. *Let f_t be a frame in a time t of a digital video with N frames. The general visual rhythm image I_{GVR} is the result of the following transformation:*

$$I_{GVR}(t,z) = \tau(f_t, z),$$

where $z \in [0, L-1]$ (L depends on the transformation function τ) and $t \in [0, N-1]$, where N corresponds to the width of the visual rhythm image.

As can be seen from Definition 2.1, only local information of a frame can be preserved when the visual rhythm for video content simplification is used. For broadcast soccer videos, it is interesting to take into account more global information on the frames, mainly if it is necessary to identify the playground color. For such purpose, the following statistical mode (Definition 2.3) is used as a transformation function. This function will be considered in a new visual rhythm representation (Definition 2.4) which deals with values of the frames histogram after color reduction procedure mentioned in Section 2.1.

Definition 2.3 Mode. *A mode is the most common element of a set of samples. Let $Hist[x]$ be the frequency of x for a numerical data sample S with P elements. The mode M is defined as follows:*

$$M(S) = m, \forall x \in [0, P-1], Hist[m] >= Hist[x].$$

In other words, the mode is the value for which the histogram reaches a maximum.

Definition 2.4 Mode Visual Rhythm. *Let $I_{GVR}(t,z)$ be a general visual rhythm as in Definition 2.2, and $f_t(z)$ a line z from a frame in time t. The mode visual rhythm image, I_{MVR}, is defined by considering the following transformation function:*

$$I_{MVR} = \tau(t,z) = M(f_t(z)),$$

where $z \in [0, H-1]$ and $t \in [0, N-1]$, H and N are, respectively, the height and the width of the mode visual rhythm image. H indicates also the height of the video frames.

Since the playground information is present in most of the soccer video shots, the mode visual rhythm (Definition 2.4) alone cannot, for example, discriminate frames containing several players from others showing just one of them (e.g., in a zoom camera work). Furthermore, it is important to know if the playground color yields a representative mode, i.e., if it has few different colors by line indicating that this mode is indeed representative. For this purpose, we also compute the mode rate visual rhythm given by the following definition.

Definition 2.5 Mode Rate Visual Rhythm. *Let $I_{GVR}(t,z)$ be a general visual rhythm as in Definition 2.2, and $f_t(z)$ a line z from a frame in time t.*

The mode rate visual rhythm, I_{MRVR}, is the result of the following transformation function:

$$I_{MRVR} = \tau(t,z) = 100 \times \frac{Hist_z[M(f_t(z))]}{W},$$

where, again, $z \in [0, H-1]$ and $t \in [0, N-1]$, H and N are, respectively, the height and width of the mode rate visual rhythm image. As before, H and W corresponds also to the height and the width of the video frames.

Figure 2 illustrates the definition of the mode visual rhythm (MVR) and mode rate visual rhythm (MRVR) representations. As we can see from this Figure, the color reduction proposed here constitutes the first step in the definition of these images containing both local and global information about the whole video. For example, besides the detection of the dominant color of the sequence considered in Figure 2, one can notice the vertical lines in the MVR and MRVR images corresponding to shot transitions. In this work, these images were used to perform key-frame extraction as discussed in the next section.

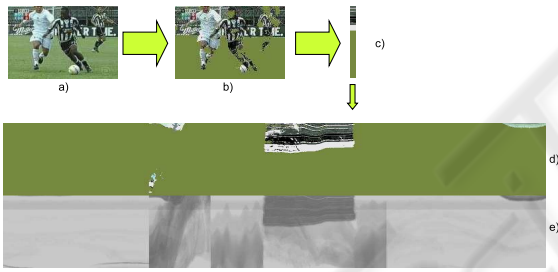

Figure 2: MVR and MRVR images of a soccer video sequence. a) original frame; b) result of color reduction; c) column with all mode by line; d) MVR image; e) MRVR image.

2.3 Key-frame Extraction Method

During broadcast TV soccer matches, different camera views are commonly applied to the whole scene. In general, most of the match is transmitted by considering a long view camera in which it is possible to have a global view with many players shown at the same time. Long view cameras are usually related to the match play time. A medium view is generally used to focus on some players actions, showing no more than two or three players at the same time. In special events, a close-up is used to show the players' face, t-shirts, referee, and also audience and coach. Close-ups are mainly related to the break intervals of the match. The next shot classification takes into account the MVR and MRVR images during a given shot interval.

2.3.1 Shot Classification

By considering the MVR and MRVR images, it is possible to identify significant shots according to some predefined rules. In the literature, different shot classification approaches are proposed aiming at different goals. (Dufaux, 2000) defines a single shot as the representative shot of a whole video. Later, he considers a key-frame selection on this shot to define one key-frame for the whole video. (Vendrig and Worring, 2003) use a shot selection to help their video annotation process by considering different classes of features, such as characters.

In this work, four classes related to the TV broadcast standard for soccer games are proposed, as illustrated in Figure 3. All cameras, for this kind of videos, use a three predefined apertures and, during the match, these apertures do not change much. Initially, the large view camera focus on a broad region of the field, related to a **class a** described further in the paper. A medium view camera records some close-ups of the players' actions, such as passing or shooting, and is associated to **classes a** and **b**. A **class c** can be related to players close-ups showing, for instance, their t-shirt, their numbers or faces. A **class d** is only considered to identify transmission errors. Basically, our shot classification approach takes into account the amount of playground color information, which, in turn, is used in our key-frame extraction method. More specifically, these four classes are:

Class a - High Dominant Green Shots: Shots with a high number of pixels related to the playground color. In general, shots corresponding to global or medium views of the field (Figure 3a).

Class b - Medium Dominant Green Shots: Shots with a medium number of pixels related to the playground color. They are commonly associated to close-up views inside the field or lateral views (Figure 3b).

Class c - Low Dominant Green Shots: Shots with a low number of pixels related to the playground color. These shots represent mainly the audience or lateral close-up views, or even other views outside the field (Figure 3c).

Class d - Non-representative Shots: In general, these shots have no significant information or without information. They are related to transmission failures or no signal segments (Figure 3d).

The classification method is based on the playground color in the MVR image and on the rate of this color indicated in the MRVR image. For each shot, the corresponding segment in the I_{MVR} and I_{MRVR} images is considered by computing the percentage of

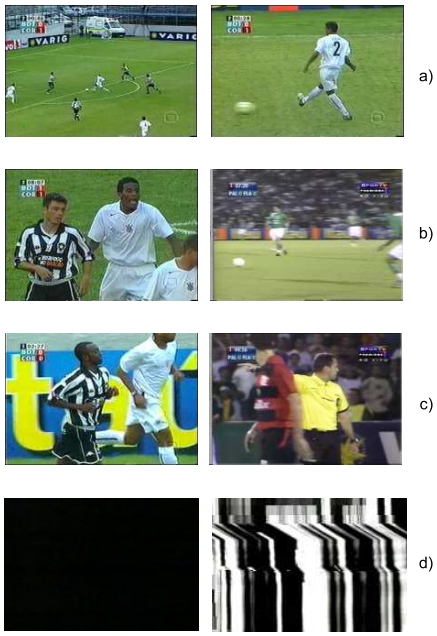

Figure 3: Example of frames belonging to the four shot classes.

the playground color and the average rate of the playground color area.

Let A be a selected shot area and PC define the set of points corresponding to the playground color in the MVR image. Also, let RA define the set of points in the MRVR image corresponding to the value of the mode rate equal or greater than 50%. This thresholding identifies all significant modes related to the MVR image and, consequently, to the background information in each line of the video frames. A basic shot classification is obtained as follows:

1. For each detected shot in the MVR image, consider the PC area. Let PCP define the PC percentage related to the shot area A, i.e., $PCP = |PC|/A$.
2. For each detected shot, consider the RA area.
3. Classify the corresponding shots using the PC and RA informations as follows:

   Class a - High Dominant Green Shots: The PCP and the percentage of the subset $PC \subset RA$ related to the shot area are equal or greater than a threshold T_1 ($\frac{|PC \subset RA|}{A} \geq T_1$).

   Class b - Medium Dominant Green Shots: The PCP is equal or greater than a threshold T_2, and the percentage of the subset $PC \subset RA$ related to the shot area is lower than the threshold T_1 ($\frac{|PC \subset RA|}{A} < T_1$).

   Class c - Low Dominant Green Shots: The PCP is lower than the threshold T_2 and the percentage of the subset $\overline{PC} \subset RA$ related to the shot area is lower than the threshold T_1 ($\frac{|\overline{PC} \subset RA|}{A} < T_1$).

   Class d - Non-representative Shots: The PC is lower than the threshold T_2 and the percentage of the subset $\overline{PC} \subset RA$ related to the shot area is equal or greater than the threshold T_1 ($\frac{|\overline{PC} \subset RA|}{A} \geq T_1$).

Figure 4 shows a video segment with eight shots. After applying a simple threshold in the MRVR image, the area corresponding to the mode rate is selected (Figure 5).

Figure 4: MVR and MRVR superimposed images with eight shots (separated by black lines).

Figure 5: Selected area from the MRVR image in Figure 4, after applying a threshold corresponding to 50% of the mode rate.

2.3.2 Key-frame selection

Once all shots are discriminated according to the previous classes, the shot length and class are finally used to determine how the key-frames are selected as our final result. In this step, the duration of the shot classified as **class a** is taken into account in order to determine the number of key-frames to be selected.

Thus, shots previously classified as belonging to **class b** and **class c** with short duration need only one key-frame to represent them. In this case, it is selected the frame located at the position corresponding to 10% of the shot length. Shots from **class a** can be of short or long duration. When long-duration shots are detected, we take into account its length and at least one key-frame at the begining (same position for the shots in **class b** and **class c**), and one at the end, corresponding to 90% of the shot length. Finally, shots belonging to **class d** have no key-frame, since these shots convey no significant information about the scene.

3 Experimental Results

Some experiments were performed using three different matches of the First Brazilian League (about 4.5 hours), from different TV channels. We consider a typical and simple shot detection approach to perform this first task with a precision of at least 82%. The simple pixel-wise comparison was used, frame by frame, as discussed by (Brunelli et al., 1999).

The color subquantization approach presented very good reduction property without losing color information and properly selecting the playground color in all three video sequences. It is important to highlight that one video has players with green t-shirts, and all of them were recorded at different times and weather conditions.

Table 1 shows the results concerning the shot classification step in which all classes were correctly classified. The threshold T_1 was defined as 82% and T_2 as 18% taking into account the three soccer videos used in the experiments. Shortly, the main task here is to classify shots according to the different playground occurrence and those representing transmission failures. Using this shot classification, the next step is to apply the key-frame selection as discussed in Section 2.3.2.

Table 1: Shot classification results.

Classes	a	b	c	d
a	934	0	0	0
b	0	887	0	0
c	0	0	621	0
d	0	0	0	1

Correctly Classified	2443	100.0%
Falsely Classified	0	0.0%
Class a	934	38.23%
Class b	887	36.31%
Class c	621	25.42%
Class d	1	0.04%
Total of Shots	2443	100.0%

A key-frame extraction validation is an arduous task mainly due to the many video characteristics. For soccer videos, it is important to whole keep the dynamic of the match through the set of the defined key-frames. To illustrate the results of our approach, the key-frames extracted here are compared to those obtained from the IBM Multimedia Analysis and Retrieval System - Marvel Lite 3.2a (Smith et al., 2007). Table 2 presents the results of the proposed approach and the IBM Marvel Lite software with respect to the number of defined key-frames.

An example of detected key-frames is shown in Figure 6 for a given video segment. Note that all the obtained key-frames describe different camera views or position, thus expressing the dynamic of the entire game, as expected. The IBM Marvel Lite software was used with its standard parameters and the corresponding key-frame extraction is presented in Figure 7 for the same video segment considered before. This result shows that some key-frames are not too relevant and that a frame belonging to a dissolve shot transition was included in the set of extracted key-frames. Note that the shot detection method used in the IBM system is not the same as the one considered here.

Table 2: Key-frame extraction results.

	Key-frames
Total of shots	2443
Proposal approach	2913
IBM Marvel Lite 3.2a	1949

Figure 6: An example of key-frames resulting of the extraction approach for a soccer video segment.

4 CONCLUSIONS

This work presented a new key-frame extraction approach for TV broadcast soccer videos. It also proposed an efficient color reduction which determines the dominant color of the soccer fields playground, as well as a novel video representative image and a shot classification method based on this playground information.

From the results illustrated above, we can see that the classification of the shots, based on their representative images, yielded a good classification method. For the key-frame extraction, our approach considers

Figure 7: Key-frames extracted by the IBM Marvel software for the same video segment as in Figure 6.

more than one key-frame per shot which, for navigation purposes, can highlight interesting segments of a soccer video.

As future works, we plan to use the MVR and MRVR images for shot detection by exploring its temporal feature, and improve the shot classification in order to classify shots by camera view modes and not only by playground appearance. Finally, it is also interesting to perform, as a validation scheme, a user test for video browsing and navigation by considering the key-frame extraction proposed here.

ACKNOWLEDGEMENTS

The authors are grateful to the National Council for Scientific and Technological Development (CNPq), CAPES and FAPESP for the financial support.

REFERENCES

Arman, F., Depommier, R., Hsu, A., and Chiu, M.-Y. (1994). Content-based browsing of video sequences. In *MULTIMEDIA '94: Proceedings of the second ACM international conference on Multimedia*, pages 97–103, New York, NY, USA. ACM Press.

Bezerra, F. N. and Leite, N. J. (2003). Video transition detection using string matching: preliminary results. In *SIBGRAPI XVI Brazilian Symposium on Computer Graphics and Image Procesing*, pages 339–346.

Brunelli, R., Mich, O., and Modena, C. M. (1999). A survey on the automatic indexing of video data. *Journal of Visual Communication and Image Representation*, 10(2):78–112+.

Chung, M. G., Lee, J., Kim, H., Song, S. M.-H., and Kim, W. M. (1999). Automatic video segmentation based on spatio-temporal features. *Korea Telecom Journal*, 4(1):1–13.

Ciocca, G. and Schettini, R. (2005). Dynamic key-frame extraction for video summarization. In Santini, S., Schettini, R., and Gevers, T., editors, *Internet Imaging VI*, volume 5670, pages 137–142. SPIE.

Doulamis, A. D., Doulamis, N. D., and Kollias, S. D. (2000). A fuzzy video representation for video summarization and content-based retrieval. *Signal Processing*, 80(6):1049–1067.

Dufaux, F. (2000). Key frame selection to represent a video. In *International Conference on Image Processing*, volume 2, pages 275–278.

Guimarães, S. J. F., Leite, N. J., Couprie, M., and de Albuquerque Arajo, A. (2003). Video segmentation based on 2D image analysis. *Pattern Recognition Letters*, 24(7):947–957.

Kim, H., Lee, J., Yang, J.-H., Sull, S., Kim, W. M., and Song, S. M.-H. (2001). Visual rhythm and shot verification. *Multimedia Tools and Applications*, 15(3):227–245.

Komlodi, A. and Marchionini, G. (1998). Key frame preview techniques for video browsing. In *DL '98: Proceedings of the third ACM conference on Digital libraries*, pages 118–125, New York, NY, USA. ACM Press.

Koprinska, I. and Carrato, S. (2001). Temporal video segmentation. *Signal Processing: Image Communication*, 16(5):477–500.

Liu, F., Dong, D., Miao, X., and Xue, X. (2003). A fast video clip retrieval algorithm based on va-file. In Yeung, M. M., Lienhart, R. W., and Li, C.-S., editors, *Storage and Retrieval Methods and Applications for Multimedia 2004*, volume 5307, pages 167–176. SPIE.

Ngo, C. W., Pong, T. C., and Chin, R. T. (1998). Survey of video parsing and image indexing techniques in compressed domain. *Symposium on Image, Speech, Signal Processing, and Robotics (Workshop on Computer Vision)*, 1:231–236.

Pardo, A. (2006). Pixel-wise histograms for visual segment description and applications. In *CIARP2006, Lecture Notes in Computer Science*, volume 4225/2006, pages 873–882. Springer.

Patel, N. V. and Sethi, I. K. (1996). Compressed video processing for cut detection. *Visual Image Signal Processing*, 143(5):315–323.

Simões, N. C. (2004). Detecão de algumas transiões abruptas em segncias de imagens (in portuguese). Master's thesis, Institute of Computing - UNICAMP.

Smith, J. R., Natsev, A. P., Tesic, J., Lexing Xie, R. Y., Letz, F., Penz, C., Seidl, J., and Yang, J. (2007). IBM multimedia analysis and retrieval system - Marvel Lite 3.2a. *http://www.alphaworks.ibm.com/tech/imars*.

Sze, K.-W., Lam, K.-M., and Qiu, G. (2005). A new key frame representation for video segment retrieval. *Circuits and Systems for Video Technology, IEEE Transactions on*, 15(9):1148–1155.

Tse, T., Marchionini, G., Ding, W., Slaughter, L., and Komlodi, A. (1998). Dynamic key frame presentation techniques for augmenting video browsing. In *AVI '98: Proceedings of the working conference on Advanced visual interfaces*, pages 185–194, New York, NY, USA. ACM Press.

Ueda, H., Miyatake, T., and Yoshizawa, S. (1991). Impact: an interactive natural-motion-picture dedicated multimedia authoring system. In *CHI '91: Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 343–350, New York, NY, USA. ACM Press.

Vendrig, J. and Worring, M. (2003). Interactive adaptive movie annotation. *IEEE MultiMedia*, 10(3):30–37.

Wolf, W. (1996). Key frame selection by motion analysis. In *IEEE International Conference on Acoustics, Speech, and Signal Processing*.

Xiong, W., Lee, J. C.-M., and Ma, R.-H. (1997). Automatic video data structuring through shot partitioning and key-frame computing. *Mach. Vision Appl.*, 10(2):51–65.

Yeung, M. M. and Liu, B. (1995). Efficient matching and clustering of video shots. In *ICIP '95: Proceedings of the 1995 International Conference on Image Processing (Vol. 1)-Volume 1*, page 338, Washington, DC, USA. IEEE Computer Society.

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

Zhang, H. J., Low, C. Y., Smoliar, S. W., and Wu, J. H. (1995). Video parsing, retrieval and browsing: an integrated and content-based solution. In *MULTIMEDIA '95: Proceedings of the third ACM international conference on Multimedia*, pages 15–24, New York, NY, USA. ACM Press.

Zhong, D., Zhang, H., and Chang, S.-F. (1996). Clustering methods for video browsing and annotation. In Sethi, I. K. and Jain, R. C., editors, *Storage and Retrieval for Still Image and Video Databases IV*, volume 2670, pages 239–246. SPIE.