Shot type characterization in 2D and 3D video content

Ioannis Tsingalis, Anastasios Tefas, Nikos Nikolaidis and Ioannis Pitas

Department of Informatics, Aristotle University of Thessaloniki
Greece
{tefas,nikolaid,pitas}@aiia.csd.auth.gr

***Abstract*—Due to the enormous increase of video and image content on the web the last decades, automatic video annotation became a necessity. The successful annotation of video and image content facilitate a successful indexing and retrieval in search databases. In this work we study a variety of possible shot type characterizations that can be assigned in a single video frame or still image. Possible ways to propagate these characterizations to a video segment (or to an entire shot) are also discussed. Finally, in the case of 3D (stereo) video, the disparity information is used to detect certain shot types (e.g. over the shoulder ones).**

I. Introduction

Because of the tremendous increase of image and video content on the web, efficient automatic video annotation became necessary for better archival, indexing and retrieval. For that reason a variety of annotation tools has been developed [1]. Moreover, with the breakthrough of 3D cinema and 3DTV, 3D video and image content increase even more the existing video/image content and the annotation types.

In this work we discuss possible characterizations of individual still images, video frames, video segments (typically video shots) and specific regions of interest (ROIs) within a video frame/image. The latter are usually by a rectangular bounding box and contain object of interest. All characterizations may use 2D or 3D information. In the case of stereo video, disparity maps extracted from the two channels provide useful additional information to characterize video frames, segments and shots.

A typical object of interest in a movie is the actor's face that can be extracted by applying appropriate face detection and tracking algorithms [2] [3] [4] [5]. Based on the characterization of this bounding box, one can annotate the entire video frame by applying a propagation rule (see Section V). In region-based characterizations, the framed object can be characterized in terms of its geometric position, motion behaviour, interactions with other objects, importance (or saliency) and dominance into the frame.

In sports videos, shot type characterizations are directly assigned into the entire video frame and not to a specific region of interest as usually happens in movies. Features like the grass-ratio, which is the ratio of the apparent grass area to the total frame area and describe the general content of a video frame, can be used to derive such characterizations.

In both cases, we may finally obtain a video frame characterization. Such characterizations can be propagated into an entire video segment or even to a video shot. The contributions of this paper are as follows:

- Region or frame-based characterizations are discussed.
- A novel method for Over-the-Shoulder (OTS) shot type detection based on stereoscopic information is presented.
- Rules that can be used in order to propagate the derived annotations from ROI level to frame level or to a video segment (shot) level are presented.

The paper is organized as follows: In Section II basic annotations are mentioned. The proposed OTS detection method is detailed in Section III, whereas its experimental evaluation is described in Section IV. Methods that can be used to propagate annotations form ROIs to frames or to shots are described in Section V. Conclusions, are drawn in Section VI.

II. Annotation Types

In this section basic video units like the ROI, Moving ROI, video frame, and video segment (typically a video shot) are considered. More specifically, we discuss about characterizations that are assigned in these units. Moreover, 2D and stereo video content is used.

We start with shot type characterizations that are assigned directly into an entire frame. As already mentioned these characterizations are typically used in sports videos [6]–[14]. Features like MPEG motion vectors [6], [10], the grass-ratio [9], [11], [12] and color histograms [13], [14] have been applied. Because sports video content has certain particularities only a limited number of shot type characterizations exist, such as *Close Up*, *Medium*, *Full Court* (the entire field is visible) or *Out of the Field* (only spectators are visible).

In movies, because of the more complex structure of the video content, a variety of shot type characterizations can be derived. In such content, frame-based characterizations can either be extracted from global characteristics of the video frame, e.g., optical flow, or from local (ROI) characteristics. Characterization/classification methods methods based on motion characteristics, extracted by applying structure tensor analysis [15], saliency maps [16] [17], optical flow [16], geometric information of the scene [16] [17], texture gradients [18] and face-related geometric information [19] [20] have

2014 IEEE 16th International Workshop on Multimedia Signal Processing (MMSP), Sep. 22-24, 2014, Jakarta, Indonesia.
978-1-4799-5896-2/14/$31.00

been proposed. Frequently used shot type characterizations of movie content are *eXtreme Long Shot (XLS)*, *Long Shot (LS)*, *Medium Long Shot (MLS)*, *Medium Shot (MS)*, *Medium Close Up (MCU)*, *Close Up (CU)*, *eXtreme Close Up (XCU)*. A detailed description of these types can be found in [20]–[23]. Other characterizations, like *Over-the-shoulder (OTS)* (Figure 2) are more difficult to be derived.

Moreover, when ROI information is used, a variety of characterizations can result. Position characterizations of the ROI, i.e., the ROI is located on the left side, on the right side, on the center, on top right of the frame, e.t.c, can be obtained. Moreover, using stereoscopic information one can decide how far or close to the viewer the object of interest within the ROI is. This can be done by calculating the mean disparity of the pixels contained in the ROI. Such object ROI characterizations complement the ones mention above, that are restricted in the 2D video content.

In addition, in stereoscopic content characterizations regarding the position of the object with respect to the screen (in front, behind or on the screen) can be derived. Using such characterizations in movies, when the object of interest is an actor's face, one can obtain information regarding the tension of the shot. For example, faces that are in front of the screen or perform a pop-up motion, usually signify a shot with significant tension. In still images, a pop-up effect refers to the case when a person or object appears to be standing up or popping out from the rest of the frame. Moreover, a moving ROI, i.e., a ROI moving over time, is characterized as pop-up, if it starts its motion form a position behind the screen and finishes its movement in front of the screen. In order to characterize a moving ROI as performing a pop-up motion we can study the 3D content of the moving region. More specifically, we can calculate the mean disparity of the ROI. A ROI which starts with a mean disparity value larger than or equal to zero (behind or on the screen) and ends with a mean disparity value smaller than zero (in front of the screen) can be characterized as performing a pop-up motion.

The relative motion of two objects of interest can also be considered. More specifically, the objects of interest may be moving away or approach each other and such information can be used to annotate the corresponding ROIs. By deriving such characterizations, one can extract high level concepts/annotations especially in movie video content, where the objects of interest are actor faces.

III. OVER THE SHOULDER IDENTIFICATION

In a OTS shot, the camera is placed behind the shoulder of an actor and captures whatever it is pointing at, usually an other actor. This configuration is used to place the viewer in the first actor's perspective and usually employed in shots with tension. Over the shoulder identification as a shot type is a difficult problem when working only with color/grayscale image features. Here we exploit the disparity information, that is implicitly available in stereo video, to solve this problem. For the disparity map estimation the algorithm in [24] was applied.

Let $\mathbf{D}$ be the disparity map of dimensions $W \times H$. We subdivide $\mathbf{D}$ into not-overlapping patches, as shown in Figure 1. Let $\mathbf{P}_i$, $i = 1, \ldots, \frac{HW}{rc}$, be an $r \times c$ pixels sub-matrix corresponding to each patch. We vectorize each $\mathbf{P}_i$ into a column vector of length rc called $\mathbf{v}_i$. The disparity patches are traversed in column major order. We compile all $\mathbf{v}_i$ into matrix $\mathbf{V}$, the i-th column of $\mathbf{V}$ being the vector $\mathbf{V}_i$.

$$\mathbf{V} = [\mathbf{v}_1 \ldots \mathbf{v}_{\frac{WH}{rc}}]$$

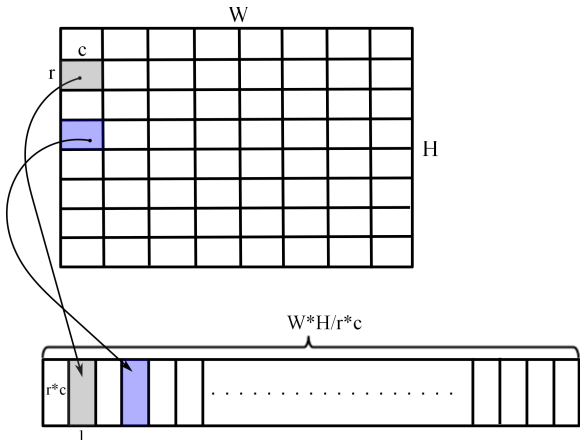

Fig. 1. Patch map creation

The facial image ROI of the second actor (the actor seen over the shoulder of the first actor) is assumed to be known and its coordinates are stored in the vector $\mathbf{c}_{roi}$. The histogram of the $\mathbf{ROI}_{face}$ is computed and the dominant disparity value, i.e, the most frequent disparity value d_{ROI} is found.

The difference, $d_{ROI} - d_{patch}$, between the dominant disparity value of the region that frames the face and the dominant disparity value of the running patch is computed and stored in a feature vector map, $\mathbf{F_{map}}(i), i = 1, \ldots, \frac{HW}{rc}$. The computed disparity differences are normalized in the range [-1, 1] with the hyperbolic tangent function:

$$f_n(x) = \frac{(1 - e^{-0.7x})}{(1 + e^{-0.7x})}.$$

Algorithm 1 Feature map extraction

Input: $\mathbf{D}$, $\mathbf{c}_{roi}$
Output: $\mathbf{F}_{map}$

$\mathbf{D} \Rightarrow \mathbf{V}$ ▷ *Obtain patches*
$\mathbf{ROI}_{face}$ = EXTRACTROI($\mathbf{D}$, $\mathbf{c}_{roi}$)
$\mathbf{h}_{ROI}$ = HISTOGRAM($\mathbf{ROI}_{face}$)
$d_{ROI} = \underset{j}{\text{argmax}}\ \mathbf{h}_{ROI}(j)$
for $i = 1 \rightarrow \frac{HW}{rc}$ **do** ▷ *For each patch*
  $\mathbf{h}_{patch}$ = HISTOGRAM($\mathbf{V}_i$)
  $d_{patch} = \underset{k}{\text{argmax}}\ \mathbf{h}_{patch}(k)$
  $\mathbf{F_{map}}(i) = f_n(d_{ROI} - d_{patch})$
Return: $\mathbf{F_{map}}$

Patches whose content is located in front of the face, in the disparity map field, produce $\mathbf{F_{map}}(i)$ values greater than zero and smaller than one, whereas, patches whose content is located behind the face in the disparity map generate $\mathbf{F_{map}}(i)$ values smaller than zero and greater than minus one. Patches that are at the same depth level with that of the face contain $\mathbf{F_{map}}(i)$ values close to zero. Such patches usually overlap with the facial ROI. The constructed feature vectors, $\mathbf{F_{map}}$ were fed into a properly two class support vector machine that classified the key-frame as being OTS or not.

IV. Experimental evaluation

A dataset of 886 stereo key frames was collected. The key frames were extracted from different types of shots, i.e, XCU, CU, MCU, e.t.c. 30% of the keyframes are characterized as over the shoulder. Each keyframe, with dimensions 540×960 pixels, is divided into patches. More specifically, we worked with 20×40 pixel (r×c) patches. Using this tessellation a 27×24 grid of assigned $\mathbf{F}_{map}(i)$, $i = 1, \ldots, 648$ values was obtained.

Based on the implementation details above, a data matrix $\mathbf{D}$ of 886×648 dimensions was obtained. In other words, we have 886 samples each one consisting of 648 dimensions. This dataset was fed to the Support Vector Machine (SVM) classifier.

Generally, the effectiveness of SVM depends on the selection of the kernel, the kernel parameter γ, and the soft margin parameter C. The Radial Basis Function (RBF) kernel was used in our experiments:

$$K(x_i, x_j) = \exp(-\gamma \|x_i - x_j\|^2), \gamma > 0. \quad (1)$$

In order to find the best C and γ, grid search to a range of these parameters was applied. Five-fold cross validation was used for measuring classification performance. In each fold 80% of the frames were used for training and the remaining 20% for testing.

The mean accuracy for all classes is 91.65%. Finally, because the two classes are unbalanced the classification accuracy of each class separately was studied. The accuracy results for the not OTS and OTS is 93.62% and 86.54% correspondingly. We observe that over the shoulder classification results are satisfactory, 86.54%, even for the small number of them, 30% of the entire database.

In Figure 2, example key frames with the disparity and feature maps are depicted. The scene parts of the scene that have the same disparity level as that of the face are portrayed in gray in the feature maps (values close to zero), whereas, parts of the scene that are in smaller or bigger disparity level from that of the face are depicted in white and black color respectively. Images in rows 2-4 depict cases that could have been classified wrongly because of the 3D scene structure, i.e., objects (e.g. book) are placed between the face and the camera. Such cases are challenging since their feature maps are similar to those encountered in OTS cases. Nevertheless, the classifier learned the structure of the shoulder silhouette and classified correctly such possibly misleading cases.

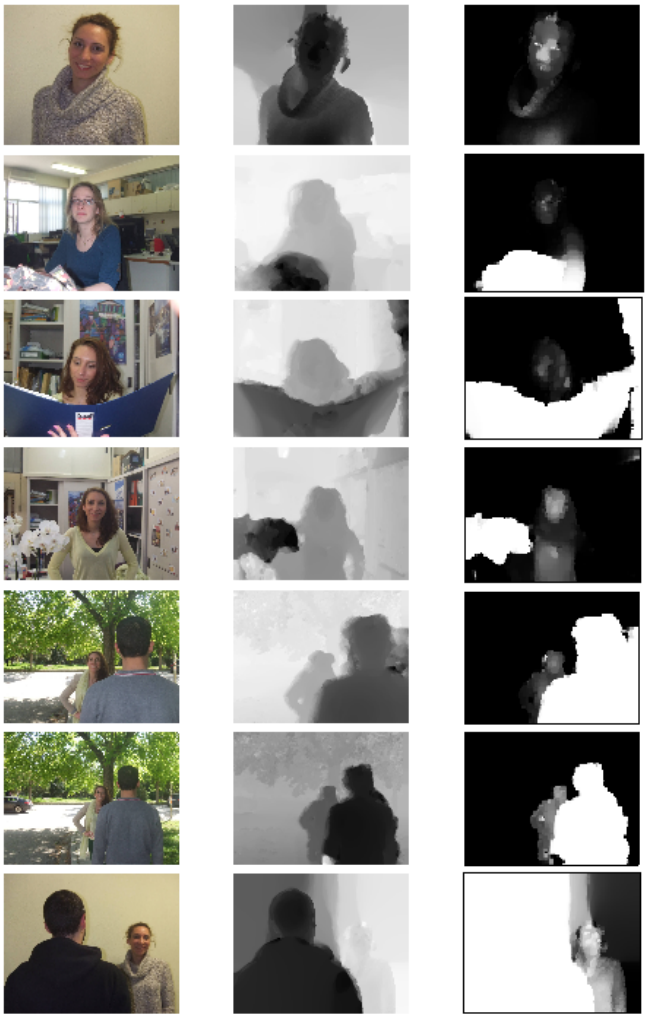

Fig. 2. Keyframes (left column), disparity maps (middle column) and the corresponding feature maps (right column). Rows 5-7: OTS. Rows 1-4: not OTS. Rows 2-4: challenging cases.

V. Propagation of annotations

As discussed in Section II, these characterizations are typically assigned either to specific ROIs in a frame or to an entire frame directly. In certain cases however there is a need to propagate these characterizations from ROIs to frames of from frames to shots or video segments. The way to perform such a propagation is not always trivial. For example, multiple ROIs may reside in a frame, each one with a specific characterization. In this case there are several possible ways to decide which type characterization will propagate to the entire video frame. Moreover, there are several possible ways to propagate a characterization from an individual video frame into an entire video shot. In this section we will try to summarize the various propagation ways.

We start with the case where a single video frame has numerous ROIs each one represented by its bounding box. Sometimes, such characterizations may contradict if they are naively propagated at frame of shot level. For example, one

person in a ROI may be characterized as depicted in a close-up, while another one may be characterized as being depicted in medium shot, depending on the way they appear in video frame. One of these contradicting characterizations should be propagated at frame and shot level. In order to assign a single characterization to the entire video frame one can use the characterization of the most dominant object ROI in the frame. The semantic dominance (or saliency) can be defined in different ways. The simplest way is to choose as dominant the one whose bounding box covers the biggest area in the video frame. Another way to make such a decision is to use the rule of thirds. According to this rule, a still-image/frame is divided into nine equally sized rectangular regions, by using two imaginary vertical and horizontal lines that divide the frame in three equally sized columns and rows. This technique is used to help photographer or cinematographer to place the objects of main interest not exactly in the center of the still-image/frame but rather in the intersection points of the vertical and horizontal lines. Using this rule, the main/dominant object, whose characterization will be propagated to the entire frame, is the one placed close to the center of the frame possibly intersected by the above-mentioned lines. Moreover, the main object of interest can be the one whose bounding box has the most salient content. A salient region semantically stands out from the rest of the image. Describing how much salient a region is can be done by applying saliency map extraction methods [25]. The output of such methods is an image whose pixel values indicate the importance of the corresponding input image pixel. By taking the mean saliency value within each ROI we can rank ROIS according to their importance. Alternatively, we can evaluate the importance of a ROI by comparing its low-level (e.g, color, depth) characteristics with the low-level characteristics of its neighbouring ROIs. In [26], a cost function is suggested for ranking a sequence of ROIs with respect to their saliency. Similarly, to the aforementioned methods that are based on 2D video features, saliency map computation using stereoscopic information is also possible [27]. By taking again the mean saliency value of each ROI, a ranking of them that describes how much an object stands out from the rest of the objects in depth is possible.

Usually, in video summarization, a video segment or shot is represented by a key-frame [28]. Having the characterization of the key-frame, derived by applying one of the aforementioned rules to the ROIs included in the key-frame, one can propagate its characterization to the entire shot. Alternatively, one can extract a single characterization for each frame based on the above and then apply majority voting to get a single characterization for the entire video segment. However, using the key-frame approach to characterize an entire shot we reduce the overall computational complexity.

A moving region (characterizing an object trajectory) is a sequence of ROIs that frames a single object in a continuous sequence of frames. ROI characterizations can be propagated at the moving region level. In order to assign a single type of characterization to a moving region, majority voting can be applied on the ROIs that form the moving region. Characterizations can also be propagated from moving regions to shots. Usually, the main object of interest has the longest trajectory in a video segment or shot. In case of multiple moving regions one can use the characterization of the longest trajectory to annotate the entire video segment or shot it belongs to.

A representative application where the aforementioned rules were applied is described in [19]. More specifically, in [19] in order to assign a characterization in shot level, a unique characterization was extracted for each frame and by applying majority voting a final shot characterization was extracted.

VI. Conclusion

In this work we summarize the possible frame and shot based characterizations in 2D and 3D video content. Moreover, possible ways to propagate the characterizations from a specific region of the image to an entire video segment are discussed. In addition, we introduce a promising way to classify video frames as over the shoulder ones. Our method is based on the 3D information contained in the disparity map. According to our knowledge there is no other work that deals with this problem using 3D content. In the future, we will try to incorporate both 3D and 2D features and train/test our model in a bigger dataset.

Acknowledgment

The research leading to these results has received funding from the European Union Seventh Framework Programme (FP7/2007-2013) under grant agreement number 287674 (3DTVS). This publication reflects only the author's views. The European Union is not liable for any use that may be made of the information contained therein.

References

[1] S. Dasiopoulou, E. Giannakidou, G. Litos, P. Malasioti, and Y. Kompatsiaris, "A survey of semantic image and video annotation tools," in *Knowledge-driven Multimedia Information Extraction and Ontology Evolution*, 2011, pp. 196–239.

[2] P. Viola and M. Jones, "Robust real-time object detection," in *International Journal of Computer Vision*, 2001.

[3] G. Stamou, M. Krinidis, N. Nikolaidis, and I. Pitas, "A monocular system for automatic face detection and tracking," in *Proceedings of Visual Communications and Image Processing (VCIP)*, July 2005, pp. 12–15.

[4] H. Baltzakis, A. Argyros, M. Lourakis, and P. Trahanias, "Tracking of human hands and faces through probabilistic fusion of multiple visual cues," in *Computer Vision Systems*. Springer Berlin Heidelberg, 2008, pp. 33–42.

[5] O. Zoidi, A. Tefas, and I. Pitas, "Visual object tracking based on local steering kernels and color histograms," *in IEEE Transactions on Circuits and Systems for Video Technology*, vol. 23, no. 5, pp. 870–882, May 2013.

[6] L. Y. Duan, M. X. Q. Tian, C. S. Xu, and J. S. Jin, "A unified framework for semantic shot classification in sports video," in *IEEE Transactions on Multimedia*, 2005, pp. 1066–1083.

[7] M. C. Tien, H. T. Chen, Y. W. Chen, M. H. Hsiao, and S. Y. Lee, "Shot classification of basketball videos and its application in shooting position extraction," in *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing*, 2007, pp. I–1085–I–1088.

[8] C. Lang, D. Xu, and Y. Jiang, "Shot type classification in sports video based on visual attention," in *Proceedings of International Conference on Computational Intelligence and Natural Computing*, 2009, pp. 336–339.

[9] A. Ekin, A. M. Tekalp, and R. Mehrotra, "Shot type classification in sports video based on visual attention," in *IEEE Transactions on Image Processing*, 2003, pp. 796–807.
[10] D. H. Wang, Q. Tian, S. Gao, and W. K. Sung, "News sports video shot classification with sports play field and motion features," in *Proceedings of International Conference on Image Processing*, 2004, pp. 2247–2250.
[11] P. Xu, L. Xie, S. Chang, A. Divakaran, A. Vetro, and H. Sun, "Algorithms and system for segmentation and structure analysis in soccer video," in *Proceedings of IEEE International Conference on Multimedia and Expo (ICME)*, August 2001, pp. 721–724.
[12] S. Chen, M. Shyu, C. Zhang, L. Luo, and M. Chen, "Detection of soccer goal shots using joint multimedia features and classification rules," in *Proceedings of 4th International Workshop on Multimedia Data Mining (MDM/KDD)*, 2003, pp. 36–44.
[13] T. Xiaofeng, L. Qingshan, and L. Hanqing, "Shot classification in broadcast soccer video," in *Electronic Letters on Computer Vision and Image Analysis*, 2008.
[14] L. Wang, M. Lew, and G. Xu, "Offense based temporal segmentation for event detection in soccer video," in *Proceedings of 6th ACM SIGMM International Workshop on Multimedia information retrieval (MIR)*, 2008, pp. 259–266.
[15] S. Wang, S. Jiang, Q. Huang, and W. Gao, "Shot type identification of movie content," in *Proceedings of IEEE International Conference on Image Processing*, October 2008, pp. 2508–2511.
[16] S. S. Benini, L. Canini, and R. Leonardi, "Estimating cinematographic scene depth in movie shots." in *Proceedings of IEEE International Conference on Multimedia and Expo (ICME)*, July 2010, pp. 855–860.
[17] M. Xu, J. Wang, M. A. Hasan, X. He, C. Xu, H. Lu, and J. S. Jin, "Using context saliency for movie shot classification," in *Proceedings of 18th IEEE International Conference on Image Processing (ICIP)*, September 2011, pp. 3653–3656.
[18] B. J. Super and A. C. Bovik, "Shape from texture using local spectral moments," in *IEEE Transactions on Pattern Analysis Machine Intelligence*, April 1995, pp. 333–343.
[19] I. Tsingalis, N. Vretos, N. Nikolaidis, and I. Pitas, "Svm-based shot type classification of movie content," in *Proceedings of 9th Mediterranean Electro technical Conference*, October 2012, pp. 104–107.
[20] I. Cherif, V. Solachidis, and I. Pitas, "Shot type identification of movie content," in *Signal Processing and Its Applications, 2007. ISSPA 2007. 9th International Symposium on*, February 2007, pp. 1–4.
[21] R. Thompson and C. Bowen, *Grammar of the Shot.* Focal Press, 2009.
[22] D. Arijon and C. Bowen, *Grammar of the film language.* James Press, 1991.
[23] J. Van Sijll, *Cinematic storytelling: the 100 most powerful film conventions every filmmaker must know.* Michael Wiese Productions, 2005.
[24] H. Hirschmuller, "Stereo processing by semiglobal matching and mutual information," in *IEEE Transactions on Pattern Analysis and Machine Intelligence*, February 2008, pp. 328–341.
[25] M. M. Cheng, N. Zhang, G.-X Mitra, X. Huang, and S.-M. Hu, "Global contrast based salient region detection," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2011, pp. 20–25.
[26] J. Feng, Y. Wei, L. Tao, C. Zhang, and J. Sun, "Salient object detection by composition," in *IEEE International Conference on Computer Vision (ICCV)*, 2011, pp. 1028–1035.
[27] Y. Niu, Y. Geng, X. Li, and F. Liu, "Leveraging stereopsis for saliency analysis," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2012, pp. 16–21.
[28] P. Kelm, S. Schmiedeke, and T. Sikora, "Feature-based video key frame extraction for low quality video sequences," in *10th Workshop on Image Analysis for Multimedia Interactive Services*, May 2009.