STEREOSCOPIC VIDEO SHOT CLUSTERING INTO SEMANTIC CONCEPTS BASED ON VISUAL AND DISPARITY INFORMATION

Konstantinos Papachristou, Anastasios Tefas, Nikos Nikolaidis and Ioannis Pitas

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

ABSTRACT

In this paper, we propose a framework for clustering shots from stereoscopic videos into clusters that correspond to semantic concepts exploiting visual and disparity information. Various color, disparity and texture descriptors are applied to shot key frames for obtaining low-level representations. Self Organizing Maps are subsequently employed upon various combinations of these representations in order to determine a lattice of representative semantic concepts. Experimental results on performances and football stereoscopic videos show that the use of disparity information leads to better clustering compared to using visual information only.

***Index Terms*—** Semantic concepts, stereoscopic video, disparity, shot clustering, Self Organizing Map

1. INTRODUCTION

Clustering of video shots has been researched mainly within a video summarization and efficient video browsing context. Indeed, clustering together shots with similar visual content is one approach/step towards summarizing a video, since the visual information in similar shots can be significantly condensed, whereas dissimilar shots should most probably form distinct parts of a visual summary. Also, shot clustering has been researched in the context of segmenting a video into scenes. Since a scene is a collection of temporally consecutive shots, one approach in doing scene detection is through clustering shots into clusters, each corresponding to a scene. Some characteristic techniques related to the above areas can be found in [1, 2, 3, 4, 5, 6].

Although 3DTV, 3D cinema have witnessed an increased popularity during the last years [7], a very limited number of shot clustering techniques operating on stereoscopic or multiview videos have been presented. Specifically, a method for multi-view video summarization including a shot clustering approach was proposed in [8]. The method represents the multi-view video structure by using a spatio-temporal shot graph, clusters the shots using random walks and generates the final summary by multi-objective optimization. [9] proposes a technique for summarization of stereoscopic videos, which performs object segmentation utilizing both color and depth information. In next, feature vectors are constructed using multidimensional fuzzy classification of segment features including size, location, color and depth, and similar shots are clustered based on the generalized Lloyd-Max algorithm.

In this paper, we propose a novel framework for shot clustering on stereoscopic video content into clusters that (hopefully) correspond to semantic concepts. Clustering of video shots into clusters that correspond to semantic concepts/tags is somewhat related to the research areas described above but also has certain important differences. This is because visually similar shots not necessarily correspond to the same semantic concept/tag, whereas shot clustering for scene detection would not group together scenes that correspond to the same concept. The main aim is to utilize disparity information, through disparity-based low-level features and to check whether this additional information can provide better clustering results compared to using low-level visual information only. Such low-level representations can be generated by employing various color, disparity and texture descriptors to shot key frames [10]. Various types of video content can be handled, notably movies, performances and sports video [11]. Once clustering is performed, an annotator can then view the results and assign semantic labels/tags such as “field long-view”, “player medium-view” to those clusters that have some meaning. The results of such a procedure can be used for stereoscopic video summarization or for metadata storage and search purposes, e.g., in AVDP format [12, 13]. The proposed approach is based on deriving combinations of low-level features and then using a Self Organizing Map for the clustering. Additionally, a way is presented to address the fact that the generated features from the various descriptors have different dimensionality.

2. PROPOSED METHOD

2.1. Feature Extraction

The proposed method operates on stereoscopic videos consisting of two visual channels (left and right). Moreover, it is assumed that disparity information has been calculated and is available in the form of a disparity channel. Let $\mathcal{V}$ be a video

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 authors views. The European Union is not liable for any use that may be made of the information contained herein.

containing N shots. Each shot is represented by the visual and disparity information of a single frame, namely the key frame, selected through a key frame selection algorithm, resulting to two image sets $\mathcal{K}^f = \{\mathbf{k}_1^f, ..., \mathbf{k}_N^f\}$ and $\mathcal{K}^d = \{\mathbf{k}_1^d, ..., \mathbf{k}_N^d\}$ containing the key frames of the left channel and the corresponding disparity maps, respectively.

Various color, disparity and texture descriptors are applied to the above key frames in order to generate low-level features. Specifically, we adopt the image representation, proposed in [14], by evaluating color/disparity histogram [15], color/disparity auto-correlogram [16], color/disparity moments, Gabor wavelet [17] and wavelet transform [18] moments. Hereafter, f_i^{jk} denotes a generated feature vector, where j can get one value of H, A, M, G, and W denoting the corresponding descriptor (histogram, auto-correlogram, moments, Gabor wavelet moments and wavelet transform moments), k can be v, d or vd depending on the kind of information (visual, disparity or visual+disparity, respectively) and i denotes the size number of elements of feature vector i. In more detail:

- Color/disparity histogram: For each visual key frame $\mathbf{k}_i^f$, a 3D HSV joint histogram $\left(f_{32}^{Hv}\right)$ is generated by uniformly quantizing its H, S and V components into 8, 2 and 2 bins, respectively. In the case of disparity key frame $\mathbf{k}_i^d$, disparity is uniformly quantized into 32 bins $\left(f_{32}^{Hd}\right)$. Also, a histogram is generated using both $\mathbf{k}_i^f$ and $\mathbf{k}_i^d$ images: H, S, V components and disparity are uniformly quantized into 8, 2, 2 and 8 bins respectively to generate a 4D HSVD joint histogram $\left(f_{256}^{Hvd}\right)$.
- Color/disparity auto-correlogram: The color or disparity auto-correlogram is evaluated by quantizing separately the $\mathbf{k}_i^f$ and $\mathbf{k}_i^d$ images into $4 \times 4 \times 4$ colors in the RGB space[1], and an extra auto-correlogram is evaluated using both $\mathbf{k}_i^f$ and $\mathbf{k}_i^d$ images, quantized into $4 \times 4 \times 4 \times 4$ colors in the RGBD space. The generated feature vectors are denoted by f_{64}^{Av}, f_{64}^{Ad} and f_{256}^{Avd}, respectively.
- Color/disparity moments: Mean and standard deviation are evaluated for the R,G,B channels separately for each $\mathbf{k}_i^f$ and $\mathbf{k}_i^d$ image resulting to f_6^{Mv} and f_6^{Md} feature vectors, respectively.
- Gabor wavelet moments: Gabor wavelet filters are applied to the $\mathbf{k}_i^f$ and $\mathbf{k}_i^d$ images, spanning four scales $[0.05, 0.1, 0.2, 0.4]$ and six orientations $[0, \pi/6, \pi/3, \pi/2, 2\pi/3, \pi]$. The mean and standard deviation of the Gabor wavelet coefficients are then computed, resulting to f_{48}^{Gv} and f_{48}^{Gd} feature vectors.
- Wavelet transform moments: Wavelet transform with a 3-level decomposition is applied to the $\mathbf{k}_i^f$ and $\mathbf{k}_i^d$ images. The mean and standard deviation of the wavelet transform coefficients are then computed, resulting to f_{48}^{Wv} and f_{48}^{Wd} feature vectors.

Since the range of values of the features vectors varies widely, the features generated for all shots are rescaled in the range $[0, 1]$.

Representations of shots are then formed by various concatenations of the above feature vectors. Regarding visual features, a concatenations of all five feature vectors are used. Since a disparity map is a coarse estimation of depth and does not contain many details, applying texture descriptors, namely Gabor wavelet and wavelet transform coefficients moments, to a disparity image may be meaningless, two representations of shots based on disparity features have been tested: the first one consists of disparity histogram, auto-correlogram and moments, while the second one includes all five feature vectors. In the same way, for the representation of shots based on both visual and disparity features, sets of features that do or do not contain the texture-related features (Gabor wavelet and wavelet transform coefficients moments) are constructed. For each case (with/without texture features) two feature sets are constructed (four features sets in total). The first set contains features generated by applying the histogram and auto-correlogram descriptors on the HSVD or RGBD space and the remaining ones separately on the visual and disparity images (namely features $f_{256}^{Hvd}, f_{256}^{Avd}, f_6^{Mv}, f_6^{Md}, f_{48}^{Gv}, f_{20}^{Wv}, f_{48}^{Gd}, f_{20}^{Wd}$), while the second one includes features generated by applying all the descriptors separately on the visual and disparity images (namely features $f_{32}^{Hv}, f_{64}^{Av}, f_6^{Mv}, f_{48}^{Gv}, f_{20}^{Wv}, f_{32}^{Hd}, f_{64}^{Ad}, f_6^{Md}$, f_{48}^{Gd}, f_{20}^{Wd}). Table 1 summarises the used sets of features.

Info	Sets of Features	Abbreviation
Visual	$f_{32}^{Hv}, f_{64}^{Av}, f_6^{Mv}, f_{48}^{Gv}, f_{20}^{Wv}$	Vis
Disparity	$f_{32}^{Hd}, f_{64}^{Ad}, f_6^{Md}$	Disp1
	$f_{32}^{Hd}, f_{64}^{Ad}, f_6^{Md}, f_{48}^{Gd}, f_{20}^{Wd}$	Disp2
Visual + Disparity	$f_{256}^{Hvd}, f_{256}^{Avd}, f_6^{Mv}, f_6^{Md}$	VisDisp1
	$f_{256}^{Hvd}, f_{256}^{Avd}, f_6^{Mv}, f_6^{Md}$, $f_{48}^{Gv}, f_{20}^{Wv}, f_{48}^{Gd}, f_{20}^{Wd}$	VisDisp2
	$f_{32}^{Hv}, f_{64}^{Av}, f_6^{Mv}, f_{48}^{Gv}, f_{20}^{Wv}$, $f_{32}^{Hd}, f_{64}^{Ad}, f_6^{Md}$	VisDisp3
	$f_{32}^{Hv}, f_{64}^{Av}, f_6^{Mv}, f_{48}^{Gv}, f_{20}^{Wv}$, $f_{32}^{Hd}, f_{64}^{Ad}, f_6^{Md}, f_{48}^{Gd}, f_{20}^{Wd}$	VisDisp4

Table 1. Image Feature Sets.

2.2. Shot Clustering

After computing the features and preparing the feature sets through their concatenation, a data matrix is formed containing the shots representation. Let us denote by $\mathbf{x}_i$, $i = 1, .., N$, a vector containing the features of i-th shot, namely one of the sets summarized in Table 1. To produce clusters for grouping the shots, a 2D Self Organizing Map (SOM) [19] is used. SOM is a Neural Network consisting of a computional layer

[1] The disparity image is considered as a three-channel grayscale image (R=G=B).

with a number of neurons ($M = N_r \times N_c$) arranged in rows (N_r) and columns (N_c) where each neuron has a weight $\mathbf{w}_j$, $j = 1, .., M$. The iterative training procedure for constructing the SOM consists of the following steps:

1. Competition: For each of the input vectors $\mathbf{x}_i$, its Euclidean distance from ever SOM neuron (actually the neuron weight $\mathbf{w}_j$) is calculated. The winning neuron k (k being the neurons index) is the one with the smallest distance.
2. Cooperation: The winning neuron k determines the spatial location of a topological neighborhood h_k of excited neurons. A typical choice of h_k is the Gaussian function $h_k(n) = \exp\left(-\frac{r_{kl}^2}{2\sigma^2(n)}\right)$, where r_{kl} is the Euclidean distance between the winning neuron k and the l neighboring neuron, n is the iteration index and $\sigma(n)$ is the neighborhood size $\sigma(n) = \sigma(0)\exp\left(-\frac{n}{E}\right)$, where E is the total number of training iterations and $\sigma(0)$ is the initial neighborhood size.
3. Adaptation: Each neuron is adapted with respect to its distance from the winning neuron and the input vector $\mathbf{w}_l(n+1) = \mathbf{w}_l(n) + \eta(n)h_{kl}(n)(\mathbf{x}_i - \mathbf{w}_l(n))$, where $\eta(n)$ is the learning rate $\eta(n) = \eta(0)\exp\left(-\frac{n}{E}\right)$ and $\eta(0)$ is the initial learning rate.

2.3. Alternative Shot Representation

Shot representation presented in Subsection 2.1 consists of combinations of various feature vectors. An issue arising from such a representation is the fact that the corresponding feature vectors have different dimensionality, which may affect the contribution of each feature vector on the shot representation. For example, the color histogram and auto-correlogram are represented from 32 and 64 values, respectively, which may create bias in favor of auto-correlogram over histogram. To overcome this issue, an alternative SOM-based shot representation is created, and compared against the simple feature concatenation based representation in Subsection 2.1. More specifically, we construct a different SOM for each feature vector participating in a feature set, trained on the corresponding training data. We use the same SOM topologies (for example 4×4) for all the feature vectors. After the SOMs construction, each input vector $\mathbf{x}_i$ corresponding to a feature vector is mapped to the respective SOM by calculating its Euclidean distance from all SOM weights/neurons $\mathbf{w}_j$. The obtained distances $d_{ij} = \|\mathbf{x}_i - \mathbf{w}_j\|_2$ for all the feature types are concatenated to form the new shot representation. In the following, we refer to such representations by using the same notations as those for the features in Table 1 followed by the "SOM" suffix, e.g., Disp1SOM.

3. EXPERIMENTAL EVALUATION

In this section we present the experiments conducted in order to assess the performance of the proposed framework for shot clustering into semantic concepts in performance and sports videos, as their analysis is of big importance [11]. We have used two video datasets consisting of performances and football stereoscopic videos. The performances dataset consists of six videos depicting three concerts and three dance shows. The football dataset consists of three football matches. The disparity maps of performance and football videos were extracted using the methods described in [20, 21] and [22], respectively. Shot boundary detection and key frame selection algorithms described in [23] have been applied to the color channels of the videos in order to extract shots and a representative frame (key frame) for each shot. Table 2 lists the videos used in our experiments and the number of extracted shots for each video.

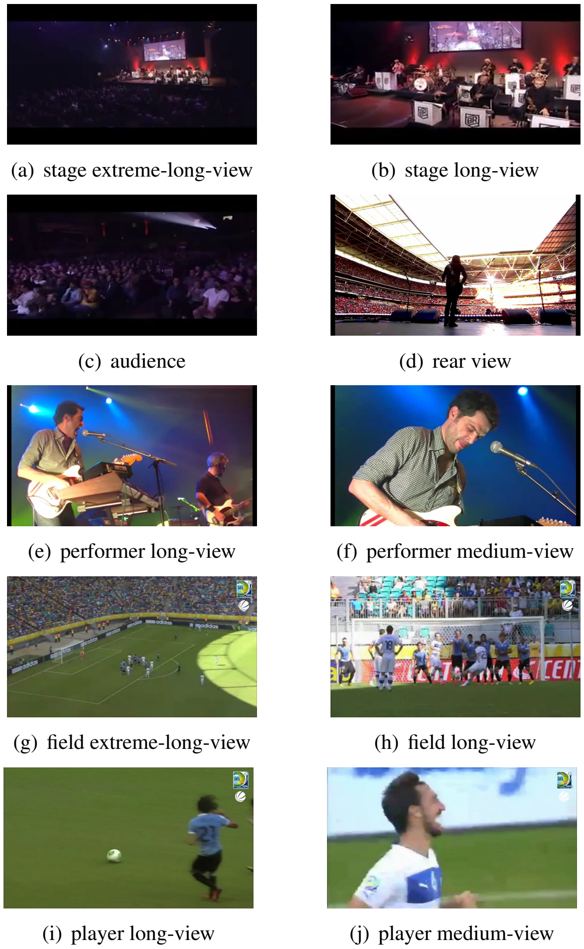
Caregory	Video Name	# shots	ground truth clusters	SOM size
Perfor-mance	1-Concert1	51	6	3×3
	2-Concert2	10	3	2×2
	3-Concert3	73	6	3×3
	4-DanceShow1	37	4	2×2
	5-DanceShow2	43	4	2×2
	6-DanceShow3	27	4	2×2
Football	7-FootballMatch1	224	4	2×2
	8-FootballMatch2	305	4	2×2
	9-FootballMatch3	223	4	2×2

Table 2. The list of videos used in our experiments.

To evaluate the various shot clustering results, we created ground-truth labels for the shots of the above videos by manually grouping the shots into a number of semantic concepts. The following labels have been used to describe these concepts: a) "stage extreme-long-view", "stage long-view", "audience", "rear view", "performer medium-view", "performer long-view" for the performances, and b) "field extreme-long-view", "field long-view", "player long-view" and "player medium-view" for the football videos. Table 2 lists the number of ground-truth clusters for each video. In Figure 1, an example image for each concept label is provided.

After extracting the features (visual, disparity, visual+disparity) as described in Subsections 2.1 and 2.3 for each video, SOMs were constructed, for each feature set of Table 1. In all the experiments, the initial learning rate $\eta(0)$, neighborhood size $\sigma(0)$ and the total number of training iterations E were set to 0.01, 8 and 350, respectively, while the number of SOM neurons for each video is selected according to the number of ground-truth clusters, as depicted in Table 2. For example, in the case of video 9, a 2×2 SOM topology was used. Finally, for the SOM constructed to generate sets of features with the same dimensionality (Subsection 2.3), a bigger initial learning rate $\eta(0)$ was selected (0.5) in order to learn the general structure of data and avoid overtraining.

Table 3 shows the best clustering performance for the visual, disparity and visual+disparity feature sets presented in



(a) stage extreme-long-view

(b) stage long-view

(c) audience

(d) rear view

(e) performer long-view

(f) performer medium-view

(g) field extreme-long-view

(h) field long-view

(i) player long-view

(j) player medium-view

Fig. 1. Examples of various semantic concepts/labels (source: Youtube).

Table 1 expressed as the mean value of the F_1 measure for 10 different random initializations of the SOM. F_1 measure is defined as $F_1 = 2\frac{PrecRec}{Prec + Rec}$, where $Prec$ and Rec denote the precision and recall measures, respectively. The corresponding best combinations of features (feature sets) are shown in brackets. As can be seen, the use of disparity information either alone or in combination with visual information leads to better performance than the use of visual information only. Specifically, the increase in the F_1 measure between the best combination of visual features and the best combination of disparity or visual+disparity features ranges from 0.020 to 0.198. Regarding the performance of disparity-related feature sets (third column in Table 3), it can be seen that sets containing texture-related descriptors (Disp2 and Disp2SOM) do not perform very well. In the case of visual+disparity feature sets, the sets VisDisp4, VisDisp3 and VisDisp4SOM usually achieve the better clustering. This means that applying the various descriptors separately to color and disparity images leads to better clustering compared to applying them to both color and disparity images (VisDisp1 and VisDisp2 cases). Additionally, the hypothesis that the different dimensionality of various feature vectors may create bias in favor of some descriptors is proven true by the fact that the SOM representations of feature sets (denoted by the "SOM" suffix) outperform the standard ones in most cases.

Video	Visual	Disparity	Visual+Disparity
1	0.376 (VisSOM)	0.380 (Disp1SOM)	**0.574** **(VisDisp4)**
2	0.485 (VisSOM)	**0.649** **(Disp1)**	0.485 (VisDisp4SOM)
3	0.724 (VisSOM)	0.633 (Disp2)	**0.746** **(VisDisp4SOM)**
4	0.362 (VisSOM)	0.401 (Disp1SOM)	**0.445** **(VisDisp4SOM)**
5	0.352 (VisSOM)	**0.501** **(Disp2SOM)**	0.400 (VisDisp1SOM)
6	0.548 (VisSOM)	0.444 (Disp1SOM)	**0.578** **(VisDisp2)**
7	0.515 (VisSOM)	0.461 (Disp1SOM)	**0.557** **(VisDisp3)**
8	0.687 (VisSOM)	0.578 (Disp1SOM)	**0.756** **(VisDisp3)**
9	0.550 (Vis)	0.474 (Disp1SOM)	**0.570** **(VisDisp4)**

Table 3. Comparative results using the F_1 measure.

Figure 2 illustrates a SOM lattice obtained for video 1 by using visual+disparity information where the closest training image to the corresponding neuron is depicted. As can be seen, disparity plays an essential role, since the top-left neurons correspond to shots where the camera is close to the subjects and thus are related to concepts such as "performer medium-view", while towards the bottom-right neurons the depth increases leading to concepts such as "stage extreme-long-view".

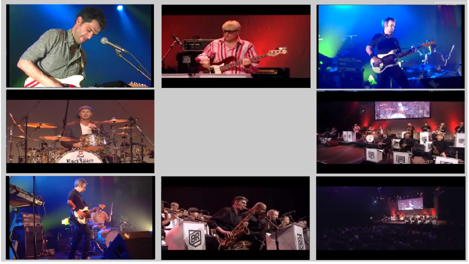


Fig. 2. A 3 × 3 SOM obtained by using visual and disparity information (source: Youtube).

4. CONCLUSIONS

In this paper, we presented a method for stereoscopic video shot clustering into semantic concepts exploiting visual and disparity information. Shots are represented by the respective key frames and various color, disparity and texture descriptors are applied to them in order to obtain low-level representations. Self Organizing Maps are constructed to obtain a topographic map of representative semantic concepts. The combination of visual and disparity features on performance and football videos achieved better clustering than the use of visual features only.

5. REFERENCES

[1] V. Chasanis, A. Likas, and N. Galatsanos, “Scene detection in videos using shot clustering and symbolic sequence segmentation,” in *9th IEEE Workshop on Multimedia Signal Processing*, 2007, pp. 187–190.

[2] Y.S. Choi, S.J. Kim, and S. Lee, “Hierarchical shot clustering for video summarization,” in *International Conference on Computational Science*, 2002, vol. 2331, pp. 1100–1107.

[3] Z. Rasheed and M. Shah, “Scene detection in hollywood movies and tv shows,” in *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2003, vol. 2, pp. 343–348.

[4] W. Tavanapong and Z. Junyu, “Shot clustering techniques for story browsing,” *IEEE Transactions on Multimedia*, vol. 6, no. 4, pp. 517–527, 2004.

[5] J. Zhang, L. Sun, S. Yang, and Y. Zhong, “Joint inter and intra shot modeling for spectral video shot clustering,” in *IEEE International Conference on Multimedia and Expo*, 2005, pp. 1362–1365.

[6] C.-W. Ngo, Y.-F. Ma, and H.-J. Zhang, “Video summarization and scene detection by graph modeling,” *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 15, no. 2, pp. 296–305, 2005.

[7] O. Schreer, P. Kauff, and T. Sikora, *3D Videocommunication: Algorithms, Concepts and Real-Time Systems in Human Centred Communication*, J. Wiley, 2006.

[8] Y. Fu, Y. Guo, Y. Zhu, F. Liu, C. Song, and Z.-H. Zhou, “Multi-view video summarization,” *IEEE Transactions on Multimedia*, vol. 12, no. 7, pp. 717–729, 2010.

[9] N. Doulamis, A. Doulamis, Y.S. Avrithis, K.S. Ntalianis, and S.D. Kollias, “Efficient summarization of stereoscopic video sequences,” *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 10, no. 4, pp. 501–517, 2000.

[10] 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*, 2009, pp. 25–28.

[11] T. D’Orazio and M. Leo, “A review of vision-based systems for soccer video analysis,” *Pattern Recognition*, vol. 43, no. 8, pp. 2911–2926, 2010.

[12] B.S. Manjunath, P. Salembier, and T. Sikora, *Introduction to MPEG-7: Multimedia Content Description Language*, J. Wiley, 2002.

[13] N. Nikolaidis, I. Pitas, K. Gorgulu, M. Liu, A. Roebel, and E.E. Tsiropoulou, “3dtv audiovisual content analysis and description,” in *11th IEEE IVMSP Workshop: 3D Image/Video Technologies and Applications*, 2013.

[14] K.-H. Yap and K. Wu, “A soft relevance framework in content-based image retrieval systems,” *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 15, no. 12, pp. 1557–1568, 2005.

[15] M.J. Swain and D.H. Ballard, “Color indexing,” *International Journal of Computer Vision*, vol. 7, no. 1, pp. 11–32, 1991.

[16] J. Huang, S.R. Kumar, and M Mitra, “Combining supervised learning with color correlograms for content-based image retrieval,” in *Proceedings ACM Multimedia*, 1997, pp. 325–334.

[17] B.S. Manjunath and W.Y. Ma, “Texture features for browsing and retrieval of image data,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 18, no. 8, pp. 837–842, 1996.

[18] J.R. Smith and S.F. Chang, “Automated binary texture feature sets for image retrieval,” in *IEEE International Conference on Acoustics, Speech, and Signal Processing*, 1996, vol. 4, pp. 2239–2242.

[19] T. Kohonen, “The self-organizing map,” *Proceedings of the IEEE*, vol. 78, no. 9, pp. 1464–1480, 1990.

[20] N. Atzpadin, P. Kauff, and O. Schreer, “Stereo analysis by hybrid recursive matching for real-time immersive video conferencing,” *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 14, no. 3, pp. 321–334, 2004.

[21] C. Riechert, F. Zilly, and P. Kauff, “Real time depth estimation using line recursive matching,” in *European Conference on Visual Media Production*, 2011.

[22] V. Kolmogorov and R. Zabih, “Computing visual correspondence with occlusions using graph cuts,” in *Eighth IEEE International Conference on Computer Vision*, 2001, vol. 2, pp. 508–515.

[23] Z. Cernekova, I. Pitas, and C. Nikou, “Information theory-based shot cut/fade detection and video summarization,” *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 16, no. 1, pp. 82–91, 2006.