2014 IEEE INTERNATIONAL WORKSHOP ON MACHINE LEARNING FOR SIGNAL PROCESSING, SEPT. 21–24, 2014, REIMS, FRANCE

STEREOSCOPIC VIDEO SHOT CLASSIFICATION BASED ON WEIGHTED LINEAR DISCRIMINANT ANALYSIS

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 stereoscopic video shot classification that includes low-level representations exploiting visual and disparity information and determination of optimal discriminant subspaces based on Linear Discriminant Analysis (LDA). Low-level representations are obtained through various color, disparity and texture descriptors which are applied to shot key frames. A new LDA-based subspace representation is proposed aiming at that optimal utilization of both visual and disparity information. The proposed shot classification framework has been evaluated on football stereoscopic videos providing enhanced classification performance and class discrimination, in comparison to using visual information only and standard LDA.

Index Terms— Shot classification, stereoscopic video, disparity, Linear Discriminant Analysis (LDA).

1. INTRODUCTION

Shot classification has been researched within a video summarization context in order to provide efficient video indexing, retrieval and browsing mechanisms in sport, news, broadcasting, etc. Video summarization methods [1, 2] initially try to select a set of salient video frames such as shot key frames, that represent the video context. Such key frames can be represented [3, 2] by color-based features including color histograms, color moments, color correlograms, etc, texture-based features such as orientation features and wavelet transformation-based texture features, and shape-based features that can describe object shapes using for example information related to the detected edges. Various clustering and classification-based techniques have been proposed to organize shot key frames into semantic groups based on Hidden Markov models [4], Hierarchical Tree structures [5], Bayesian classifier and Support Vector Machines [6], Graphs [7], Fuzzy classification [8], etc.

Although 3DTV and 3D cinema have witnessed an increased popularity during the last years [9], a very limited number of video summarization techniques operating on stereoscopic or multiview videos have been presented and are mainly using a shot clustering-based approach. Specifically, a method for multi-view video summarization was proposed in [10] which 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. A technique for summarization of stereoscopic videos was presented in [11], 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 classification of shots from stereoscopic video content to semantic classes such as “field long-view” and “player medium-view”. The main aim is to utilize disparity information and to check whether this additional information can enhance better classification results compared to using visual information only. More specifically, video shots are represented by key frames and the corresponding disparity maps, and low-level representations are generated by employing various color, disparity and texture descriptors. Classification involves a new subspace learning method based on Linear Discriminant Analysis (LDA) which takes into account the fact that key frames representations are comprised of two subsets of features (visual and disparity) in order to learn projection subspaces equipped with more generalization ability. We show that this can be achieved by adding an appropriate constraint in the objective function of LDA.

The paper is structured as follows. The proposed shot classification framework is described in Section 2. Section 3 illustrates experiment results conducted in order to evaluate its performance. Finally, conclusions are drawn in Section 4.

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

2. PROPOSED METHOD

The proposed shot classification framework operates on stereoscopic videos consisting of two visual channels (left and right) and includes two processing steps. The first one involves the representation of shot key frames using low-level features obtained by various color, disparity and texture descriptors. It is assumed that disparity information has been calculated and is available in the form of a disparity sequence and that the video has been divided into shots by using a shot cut detector algorithm. Additionally, through a key frame selection algorithm, each shot is represented by a key frame for each channel of the left channel and the disparity channel. In the second step, the obtained low-level representations are mapped to discriminant subspaces by applying a proposed extension of the Linear Discriminant Analysis (LDA) subspace learning technique and classified to the nearest class. In the following, each step of the proposed shot classification method is described in detail. The feature extraction process is presented in Subsection 2.1. The standard LDA is briefly described in Subsection 2.2. Finally, the proposed LDA extension is presented in Subsection 2.3.

2.1. Feature Extraction

Let $\mathcal{V}$ be a stereoscopic video containing N shots. Each shot is represented by the key frame of the left channel and the disparity map of the frame, resulting to two image sets $\mathcal{K}^v = \{\mathbf{k}_1^v, ..., \mathbf{k}_N^v\}$ 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. Low-level representations are generated by applying to the above key frames various color, disparity and texture descriptors adopting the image representation proposed in [12]:

- For each visual key frame $\mathbf{k}_i^v$, a 3D HSV joint histogram [13] is generated by uniformly quantizing its H, S and V components into 8, 2 and 2 bins, respectively.
- For each disparity key frame $\mathbf{k}_i^d$ a 32-bin disparity histogram is evaluated.
- The color or disparity auto-correlogram [14] is evaluated by quantizing separately the $\mathbf{k}_i^v$ and $\mathbf{k}_i^d$ images into $4 \times 4 \times 4$ colors in the RGB space[1].
- Mean and standard deviation are evaluated for the R,G,B channels separately for each $\mathbf{k}_i^v$ and $\mathbf{k}_i^d$ image.
- Gabor wavelet filters [15] spanning four scales $[0.05, 0.1, 0.2, 0.4]$ and six orientations $[0, \pi/6, \pi/3, \pi/2, 2\pi/3, \pi]$ are applied to the $\mathbf{k}_i^v$ and $\mathbf{k}_i^d$ images. The mean and standard deviation of the Gabor wavelet coefficients are then computed.
- Wavelet transform [16] with a 3-level decomposition is applied to the $\mathbf{k}_i^v$ and $\mathbf{k}_i^d$ images. The mean and standard deviation of the wavelet transform coefficients are then computed.

The list of applied descriptors is summarized in Table 1. Here, $\mathbf{f}_i^{jk}$ denotes a generated feature vector, where j can get one value of H, A, S, G, and T denoting the corresponding descriptor (histogram, auto-correlogram, moments, Gabor wavelet moments and wavelet transform moments), k can be v or d depending on the kind of information (visual or disparity, respectively) and i denotes denotes the feature dimensionality.

Descriptor	Features
HSV/Disparity Histogram	$\mathbf{f}_{32}^{Hv}$ $\mathbf{f}_{32}^{Hd}$
Auto-correlogram	$\mathbf{f}_{64}^{Av}$ $\mathbf{f}_{64}^{Ad}$
RGB moments	$\mathbf{f}_{6}^{Sv}$ $\mathbf{f}_{6}^{Sd}$
Gabor wavelet moments	$\mathbf{f}_{48}^{Gv}$ $\mathbf{f}_{48}^{Gd}$
Wavelet transform moments	$\mathbf{f}_{20}^{Tv}$ $\mathbf{f}_{20}^{Td}$

Table 1. Visual and disparity information descriptors.

Representations of shots are finally formed by the concatenation of the above feature vectors. Let $\mathcal{X} = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N\}$ be the the set of generated low-level representations $\mathbf{x}_i \in \mathcal{R}^{340 \times 1}$ of the visual ($\mathcal{K}^v$) and disparity ($\mathcal{K}^d$) key frames, where

$$\mathbf{x}_i = [\mathbf{f}_{32}^{Hd}, \mathbf{f}_{64}^{Ad}, \mathbf{f}_{6}^{Sd}, \mathbf{f}_{48}^{Gd}, \mathbf{f}_{20}^{Td}, \mathbf{f}_{32}^{Hv}, \mathbf{f}_{64}^{Av}, \mathbf{f}_{6}^{Sv}, \mathbf{f}_{48}^{Gv}, \mathbf{f}_{20}^{Tv}]_i \tag{1}$$

contains the visual and disparity information of the i-th key frame. Additionally, in our experiments we used shot representations consisting of either visual information or disparity information only. More specifically, in the case of using visual information only, the low-level representation $\mathbf{x}_i \in \mathcal{R}^{170 \times 1}$ of the i-th key frame is formed by the concatenation of visual feature vectors

$$\mathbf{x}_i = [\mathbf{f}_{32}^{Hv}, \mathbf{f}_{64}^{Av}, \mathbf{f}_{6}^{Sv}, \mathbf{f}_{48}^{Gv}, \mathbf{f}_{20}^{Tv}]_i. \tag{2}$$

Correspondingly, in the case of using disparity information only

$$\mathbf{x}_i = [\mathbf{f}_{32}^{Hd}, \mathbf{f}_{64}^{Ad}, \mathbf{f}_{6}^{Sd}, \mathbf{f}_{48}^{Gd}, \mathbf{f}_{20}^{Td}]_i. \tag{3}$$

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

2.2. Linear Discriminant Analysis

Linear Discriminant Analysis (LDA) [17] belongs to the supervised subspace learning techniques and determines a projection matrix $\mathbf{W}$ so that the classes of the projected samples

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

are well discriminated. The objective of LDA is to find the transformation matrix $\mathbf{W}$ that maximizes the following objective function:

$$J(\mathbf{W}) = \arg\max_{\mathbf{W}} \frac{\mathrm{tr}[\mathbf{W}^T \mathbf{S}_B^{LDA} \mathbf{W}]}{\mathrm{tr}[\mathbf{W}^T \mathbf{S}_W^{LDA} \mathbf{W}]}. \quad (4)$$

In the previous equation the between-class scatter matrix and the within-class scatter matrix are defined as:

$$\mathbf{S}_B^{LDA} = \sum_{i=1}^{c} \left(\boldsymbol{\mu}_i - \boldsymbol{\mu}\right)\left(\boldsymbol{\mu}_i - \boldsymbol{\mu}\right)^T \quad (5)$$

and

$$\mathbf{S}_W^{LDA} = \sum_{i=1}^{c} \sum_{k=1}^{n_i} \left(\boldsymbol{x}_k^i - \boldsymbol{\mu}_i\right)\left(\boldsymbol{x}_k^i - \boldsymbol{\mu}_i\right)^T, \quad (6)$$

where $\boldsymbol{x}_k^i$ is the k-th sample in the i-th class, $\boldsymbol{\mu}_i$, n_i are the mean vector and the number of samples in class i, respectively, and c, $\boldsymbol{\mu}$ denote the total number of classes and the mean vector of the entire dataset, respectively.

Thus, the objective of LDA is to find $\mathbf{W}$ so that in the new m-dimensional space the class means are as far from each other as possible and samples from the same class are as close to their mean as possible. The solution of (4) is approximated [18] by the following generalized eigenvalue decomposition problem:

$$\mathbf{S}_B^{LDA} \cdot \mathbf{w} = \lambda \cdot \mathbf{S}_W^{LDA} \cdot \mathbf{w} \quad (7)$$

and keeping the m eigenvectors that correspond to the m largest eigenvalues to form the columns of $\mathbf{W}$ (projection vectors). The maximum number of nonzero eigenvalues is equal to $c - 1$ and thus the upper bound on m is $c - 1$.

2.3. Weighted Linear Discriminant Analysis

Although LDA uses the data class label information for the determination of the projection matrix, it does not take into account the specific nature of data samples in certain problems. In our case, a sample of the form (3), is comprised of two subsets: the first one corresponds to disparity-related feature vectors ($\mathbf{f}_{32}^{Hd}, \mathbf{f}_{64}^{Ad}, \mathbf{f}_{6}^{Sd}, \mathbf{f}_{48}^{Gd}, \mathbf{f}_{20}^{Td}$), while the second subset contains the visual features ($\mathbf{f}_{32}^{Hv}, \mathbf{f}_{64}^{Av}, \mathbf{f}_{6}^{Sv}, \mathbf{f}_{48}^{Gv}, \mathbf{f}_{20}^{Tv}$). When we combine features with different discriminant ability, it is possible that the more discriminant will dominate and will have considerably higher projection values that the less discriminant. Moreover, if we consider a small training set that favors the one subset of features over the other, it is possible to have overtraining that will result in significantly smaller projection values for the second subset and thus their discriminant ability, even if it is smaller, will be completely lost. The proposed solution is to try regularizing the contribution of each subset to the discriminant projection by setting a-priori the contribution of each subset using a parameter b. In our case, we can assume that, ideally, the disparity features (the first K values of the sample) should hold a percentage b of the total discriminant ability, while the visual ones (the remaining $M - K$ values of the sample, where M denotes the sample dimensionality) should hold the remaining percentage. Accordingly, it would be expected that the discriminant projection vectors $\mathbf{w}$ (columns of $\mathbf{W}$) of LDA should follow a similar distribution of element values in the two groups.

To cope with the above, we modify the objective function of LDA, in order to determine projection vectors that follow a specific distribution of values defined by a weight b, so that the samples are projected in corresponding discriminant subspaces. The optimal b can be found by a line search approach. In more detail, the error of a projection vector

$$\mathbf{w} = [\underbrace{w_1, w_2, ..., w_K}_{b}, \underbrace{w_{K+1}, ..., w_M}_{1-b}]^T, \quad (8)$$

with respect to the sum of element values in the two groups defined by a weight b, can be given by the following equation:

$$E_{\mathbf{w}} = \left((1-b)\sum_{i=1}^{K} w_i - b\sum_{i=K+1}^{M} w_i\right)^2 \quad (9)$$

By expanding (9), the error can be re-written as:

$$\begin{aligned} E_{\mathbf{w}} \;&=\; (1-b)^2 \sum_{i=1}^{K} w_i^2 + b^2 \sum_{i=K+1}^{M} w_i^2 \\ &\quad -2(1-b)b \sum_{i=1}^{K} w_i \sum_{j=K+1}^{M} w_j \\ &\quad + (1-b)^2 \sum_{i=1}^{K} \sum_{\substack{j=1 \\ i \neq j}}^{K} w_i w_j \\ &\quad + b^2 \sum_{i=K+1}^{M} \sum_{\substack{j=K+1 \\ i \neq j}}^{M} w_i w_j \end{aligned} \quad (10)$$

We can notice that the first two terms correspond to the sum of the squared disparity and visual discriminant values, the next term refers to the relation of the visual and disparity discriminant values, and the two last terms refer to the relation of either the visual or the disparity discriminant values. The last two terms refer to the relation of the discriminant projection values inside the disparity subset or inside the visual subset. Thus, they can be omitted by the objective function, since we are only interested in regularizing the projection values of the one subset related to the other. The above leads to an error expression of the form:

$$\begin{aligned} E_{\mathbf{w}} \;&=\; (1-b)^2 \sum_{i=1}^{K} w_i^2 + b^2 \sum_{i=K+1}^{M} w_i^2 \\ &\quad -2(1-b)b \sum_{i=1}^{K} w_i \sum_{j=K+1}^{M} w_j \end{aligned} \quad (11)$$

Taking into account the fact that the objective function used in LDA uses a projection matrix, we define a matrix $\mathbf{A}$, whose multiplication ($\mathbf{w}^T\mathbf{A}\mathbf{w}$) with a projection vector equals to the vector error. It can be easily proven that for a $M \times M$ matrix of the form:

$$\begin{bmatrix} \begin{bmatrix} (1-b)^2 & 0 \\ 0 & (1-b)^2 \end{bmatrix}_{K\times K} & [-b(1-b)]_{M-K\times K} \\ [-b(1-b)]_{K\times M-K} & \begin{bmatrix} b^2 & 0 \\ 0 & b^2 \end{bmatrix}_{M-K\times M-K} \end{bmatrix} \tag{12}$$

$\mathbf{w}^T\mathbf{A}\mathbf{w}$ equals the error $E_\mathbf{w}$:

$$\begin{aligned} \mathbf{w}^T\mathbf{A}\mathbf{w} &= (1-b)^2\sum_{i=1}^{K} w_i^2 + b^2 \sum_{i=K+1}^{M} w_i^2 \\ &\quad -2(1-b)b\sum_{i=1}^{K} w_i \sum_{j=K+1}^{M} w_j \\ &= E_\mathbf{w}. \end{aligned} \tag{13}$$

The objective of the proposed Weighted Linear Discriminant Analysis (WLDA) is to exploit the fact that the elements of the input data follow a specific distribution for the determination of projection vectors $\mathbf{w}$, which both increase class discrimination and follow the same value distribution. To this end, we want to maximize $\mathrm{tr}[\mathbf{W}^T\mathbf{S}_B^{LDA}\mathbf{W}]$, so that the dispersion of samples from different classes will be maximized after the projection, while, at the same time, we want to minimize a) the trace of the $\mathbf{W}^T\mathbf{S}_W^{LDA}\mathbf{W}$ so that samples from the same class will come as close as possible to their mean vector after the projection and b) to minimize the trace of the $\mathbf{W}^T\mathbf{A}\mathbf{W}$, so that the error defined by (11) becomes as small as possible after the projection:

$$J(\mathbf{W}) = \arg\max_{\mathbf{W}} \frac{\mathrm{tr}[\mathbf{W}^T\mathbf{S}_B^{LDA}\mathbf{W}]}{(1-s)\,\mathrm{tr}[\mathbf{W}^T\mathbf{S}_W^{LDA}\mathbf{W}] + s\,\mathrm{tr}[\mathbf{W}^T\mathbf{A}\mathbf{W}]}. \tag{14}$$

where $s \in [0, 1]$ is a factor that controls the error of $\mathbf{w}$. It is obvious that, for $s = 0$, WLDA is equivalent to LDA, while, as $s \to 1$, the level of error of the projection vectors is minimized. The solution of (14) is approximated by the following generalized eigenvalue decomposition problem:

$$\mathbf{S}_B^{LDA}\cdot\mathbf{w} = \lambda\cdot\left((1-s)\mathbf{S}_W^{LDA} + s\mathbf{A}\right)\cdot\mathbf{w}, \tag{15}$$

by keeping the m eigenvectors that correspond to the m largest eigenvalues. The upper bound on m, as in the case of LDA, is $c-1$, where c is the total number of classes.

Figure 1 illustrates the result of LDA and WLDA on artificial (toy) data from 2 classes. Here, we assume that the horizontal and vertical axes correspond to disparity and visual information, respectively. We notice that the disparity elements correspond to 60% of the sum of visual and disparity elements. The actual distribution of data is represented by two elliptical regions, while the available samples of the two classes are represented by crosses and circles, respectively. As can be seen, the samples are not very representative of the actual distribution, while some of them can be considered as noisy data. It is clear that even if the LDA projection separates well the samples from the two classes, it does not determine a best projection for the two actual distribution. On the contrary, WLDA determines a projection line such that projection on the line best separates both the available samples and the actual distributions.

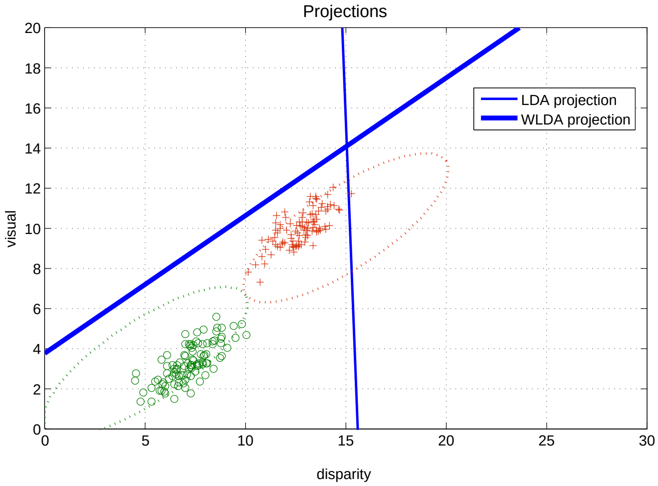
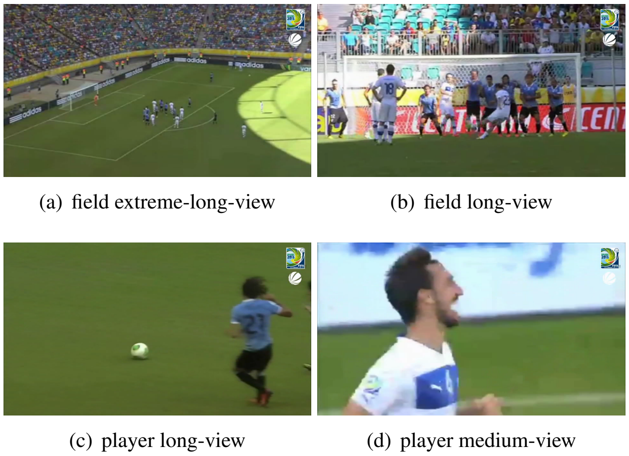


Fig. 1. LDA and WLDA projection lines for a 2-class artificial data problem.

3. EXPERIMENTAL EVALUATION

In this section we present the experiments conducted for assessing the performance of the proposed framework for shot classification. We have used a stereoscopic video dataset consisting of three soccer matches, since soccer video analysis is of special importance [19]. The disparity maps of the videos were extracted using the method described in [20]. Shot boundary detection and key frame selection algorithms described in [21] have been applied to the color channels of the videos in order to extract shots and a representative frame (key frame) for each shot. The number of extracted shots for each video were 520, 622 and 470, respectively. To evaluate the shot classification results, we created ground-truth labels for the shots of the above videos by manually grouping the shots into 4 semantic concepts/classes. The following labels have been used to describe these classes: “field extreme-long-view”, “field long-view”, “player long-view” and “player medium-view”. In Figure 2, an example image for each concept label is provided.

After extracting the features as described in Subsection 2.1 for each video, a series of experiments were conducted. In all the experiments, we applied LDA or the proposed extension (WLDA) to the training set and the samples were projected into the corresponding subspace. Additionally, LDA



(a) field extreme-long-view

(b) field long-view

(c) player long-view

(d) player medium-view

Fig. 2. Examples of the semantic concepts/labels.

was applied to the visual and disparity features separately. The projected samples were classified using the Nearest Centroid (NC) classifier. In the case of WLDA, the values of b and s were determined by using a grid search strategy. In our experiments, a combination of $b = [0.00, 0.05, 0.10, ..., 1.00]$ and $s = [0.00, 0.05, 0.10, ..., 1.00]$ was used. In our method, it is considered that a sample is comprised of two subsets (visual and disparity information) leading to searching optimal values for two parameters. Alternatively, a sample may be considered that is comprised by more subsets, based on e.g., the type of descriptor (see Table 1). In such a case, a larger number of parameters should be determined and adopting this searching procedure would be infeasible.

In the first series of experiments, we used the 5-fold cross validation procedure for each video. More specifically, the samples (key frames) for each video have been randomly spit in five non-overlapping subsets. For each video the same partitioning was used in all the experiments. In each fold, the techniques were trained by using 4 subsets and testing was performed on the remaining subset. Performance was measured by evaluating the mean classification rate over all five folds. Classification results are illustrated in Table 2. Columns 2-4 depict the classification accuracies obtained by applying the LDA on the visual, disparity and visual+disparity features, respectively. The last column depicts the recognition accuracies obtained by applying the WLDA on both visual and disparity features. The best results are shown in bold. We can observe that, in the case of LDA, using visual information only provides better classification compared to using only disparity or both visual and disparity information. Thus, it seems that disparity does not provide discriminant information. On the other hand, the proposed WLDA technique outperforms the standard LDA in all the cases providing 0.5% - 3% improvement on the performance.

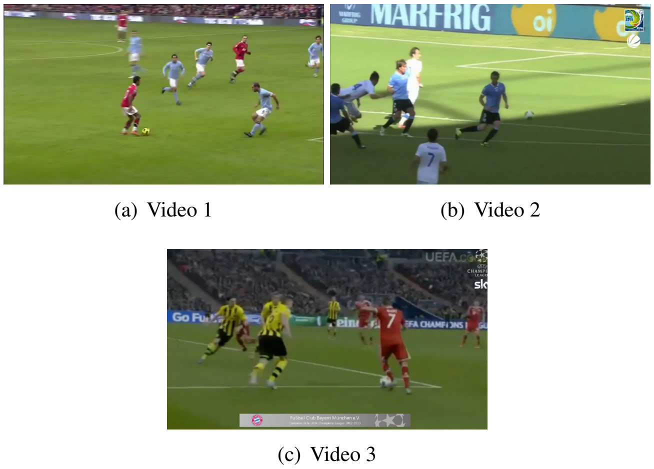
Video	LDA			WLDA
	Visual	Disparity	Both	Both
1	71.05	58.83	61.34	**72.13**
2	76.18	56.60	72.86	**76.53**
3	71.23	53.19	65.27	**73.84**

Table 2. 5-fold cross validation classification accuracies (%).

The second series of experiments included three experiments. In each experiment, the samples (key frames) of a video comprised the training set, while the test set consist of the key frames of the two remaining video. WLDA was used with the optimal values of b and s obtained by the 5-fold procedure for each video. The results obtained in these experiments are illustrated in Table 3. It can be easily observed that in all the videos the proposed WLDA provides 2.5% - 13% improvement compared to standard LDA. On the other hand, it is obvious that the performance rates are rather low in comparison with the ones obtained by the 5-fold procedure. This can be explained by the fact that the directing/filming style of the various football matches may differ. For example, Figure 3 illustrates three representative example of the class “field long-view”, one for each video, where it is obvious that different styles are used in each football match.

Video	LDA			WLDA
	Visual	Disparity	Both	Both
1	37.69	22.50	33.23	**47.73**
2	47.78	27.28	41.97	**50.11**
3	33.98	31.93	30.78	**46.96**

Table 3. Comparison of the best classification accuracies (%).



(a) Video 1

(b) Video 2

(c) Video 3

Fig. 3. Examples of class “field long-view”, for each video.

4. CONCLUSIONS

In this paper, we presented a framework for stereoscopic video shot classification exploiting visual and disparity information. Shots are represented by the respective key frames and low-level representations obtained by applying various color, disparity and texture descriptors. Optimal class representations are determined by modifying the objective functions of LDA technique. Experimental results on football stereoscopic videos showed enhanced classification performance and class discrimination compared to using visual information only and the standard LDA technique.

5. REFERENCES

[1] Ying Li, Shih-Hung Lee, Chia-Hung Yeh, and C.-C.J. Kuo, "Techniques for movie content analysis and skimming: tutorial and overview on video abstraction techniques," *IEEE Signal Processing Magazine*, vol. 23, no. 2, pp. 79–89, 2006.

[2] Weiming Hu, Nianhua Xie, Li Li, Xianglin Zeng, and S. Maybank, "A survey on visual content-based video indexing and retrieval," *IEEE Transactions on Systems, Man, and Cybernetics, Part C: Applications and Reviews*, vol. 41, no. 6, pp. 797–819, 2011.

[3] 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.

[4] Lexing Xie, Shih-Fu Chang, A. Divakaran, and Huifang Sun, "Structure analysis of soccer video with hidden markov models," in *IEEE International Conference on Acoustics, Speech, and Signal Processing*, 2002, vol. 4, pp. IV–4096–IV–4099.

[5] Jianping Fan, A.K. Elmagarmid, Xingquan Zhu, W.G. Aref, and Lide Wu, "Classview: hierarchical video shot classification, indexing, and accessing," *IEEE Transactions on Multimedia*, vol. 6, no. 1, pp. 70–86, 2004.

[6] Ling-Yu Duan, Min Xu, Qi Tian, Chang-Sheng Xu, and J.S. Jin, "A unified framework for semantic shot classification in sports video," *IEEE Transactions on Multimedia*, vol. 7, no. 6, pp. 1066–1083, 2005.

[7] 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.

[8] "A fuzzy video content representation for video summarization and content-based retrieval," *Signal Processing*, vol. 80, no. 6, pp. 1049 – 1067, 2000.

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

[10] 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.

[11] 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.

[12] 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.

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

[14] 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.

[15] 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.

[16] 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.

[17] R.O. Duda, P.E. Hart, and D.G. Stork, *Pattern Classification, 2nd ed*, John Wiley & Sons, 2001.

[18] H. Wang, S. Yan, D. Xu, X. Tang, and T. Huang, "Trace ratio vs. ratio trace for dimensionality reduction," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2007, pp. 1–8.

[19] 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.

[20] 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.

[21] 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.