Scene Selection of Broadcast Soccer Video using Temporal Slices

Masao Izumi†
†Graduate School of Engineering, Osaka Prefecture University, Japan

Abstract
In this paper, we propose a novel method for selecting personal favorite scenes using hierarchical eigenspace method based approach from broadcast soccer videos. In our method, we use physical parameters extracted from videos, such as image space frequency and time space frequency. We use these physical parameters as feature vectors, make an eigenspace of these vectors extracted from image slices of broadcasted soccer videos, and select the image slices that the distance from prior selected image slices as a personal favorite scene in the eigenspace is small. And we construct the hierarchical image pyramid which includes smaller resized images of input frames. Firstly we use the eigenspace of smallest resized frames of videos, and extract candidate frames of personal favorite scenes. In the next steps, we use the eigenspaces of next smallest resized frames. And finally we extract frames of personal favorite scenes using eigenspace of non-resized frames. The experimental results show the ability of our proposed method.

Keywords: Eigenspace method, camera works, temporal slices, frequency domain.

1 Introduction
In recent years, the spread of cable TV, DVD or Blu-ray recorders, etc., enabled individuals to record a lot of TV programs easily. But, it needs immense time and efforts to search scenes wanting to watch in large amount of videos. Then, the technology which gives an effective index automatically will be more indispensable from now on.

However, it is still difficult to create a general technique of indexing for all kinds of video. So there are many challenging researches about scene estimation, event estimation, and indexing. Leonardi at al. [1] estimate major soccer scenes using cameraworks peculiar to soccer videos, for example it pans and zooms rapidly on shoot scenes and corner kick scenes. Xinghua et al. [2] estimate goal events using textures and score boards peculiar to soccer videos. Moreover, it takes into consideration that a video is generally multiple streams of media information, such as audios, texts, and images, it will be thought that the performance of indexing can be raised more by unifying multiple media information. Uegaki et al. [3] proposed multimodal indexing from broadcast soccer video using Dynamic Bayesian networks which input cameraworks, players and ball trajectories, and audio power spectrum. These advanced researches show good performance on automatic indexing of broadcast soccer video, but these indexes are quite general explanations of scene features, such as shoot scene, free-kick scene, throw-in scene, etc. On the other hand, when these methods are implemented on the personal used equipment, such as DVD recorders, users’ demands are more personal. So sometimes the general explanation of index extracted from above methods isn’t appropriate to these personal favorite demands, because the personal demands are wide-ranged various.

From these considerations, we try to establish the method of personal favorite scene selection from broadcast soccer video in this paper. We think personal demands are so various that physical parameters such as image space frequency and time space frequency extracted from videos tend to be unique, then we propose the method of selecting personal favorite scenes using image space frequency and time space frequency.

2 Overview of Proposed Method
In the case of sports video retrieval, it is an important issue how to represent the trend of personal favorite scenes. For example, some persons are prefer shooting scenes and placement kick scenes in broadcast soccer video, the others like skillful passing scenes or powerful defense scenes. Recent researches of automatic retrieval
for sports videos are mainly focused on the search of scenes which are generally demanded by many people. But there are few researches which treat the way of searching personal favorite scenes. In this paper, we introduce the method of automatic retrieval for searching personal favorite scenes from broadcast soccer video using principal component analysis (PCA).

In this work, we focus on the scenes of ‘Playing Field Shots’. In general, there are three categories of shots in broadcast soccer videos, ‘Playing Field shots’, ‘Players shots’, and ‘Audience shots’. ‘Playing Field shot’ is a global view of soccer playing field in which many players are viewed. These shots are occupied 50-60% in general broadcast soccer videos, and viewers can recognize how the game goes in these shots. Followings are our approaches for selecting personal favorite scenes from broadcast soccer videos. Firstly, we construct the image pyramid of input video frames. This pyramid is constructed with resized images which are resized as 1, 21, 22, …, times of original input frame resolution. Secondly, we extract physical parameters such as image space frequency and time space frequency as a feature vector from smallest resized each frame of video sequence. And, we construct the eigenspace of these feature vectors as the first feature space. Next we calculate the distance from points that are projected by the feature vectors of previously selected as the personal favorite scenes. The frames which distance is below the threshold are extracted as the first candidate frames for searching personal favorite scenes. In the next step, we use the eigenspace of second smallest resized frames for selecting second candidates of personal favorite scenes. After repeating this process several times, finally we extract the final results of personal favorite scenes using the eigenspace which is calculated with the non-resized frames.

3 Feature Vectors

Feature vectors should include not only the feature of each frame, such as color histograms and image space frequency which represent the feature of instant at the time of each frame. But also the feature vectors should include the feature of time sequential changes between the time of each frame and the time of several frames later. In this research, we use space frequency of pixel intensities for the representation of the features at the time of each frame, and time space frequency of pixel intensities on x-t and y-t planes of time sequential images for the representation of the features between the time of each frame and several frames later. Feature vector \( v(t) \) is defined as follows. This definition is for the original size frames of video sequence. Feature vectors of other resized frames are defined as the same manner.

\[
v(t) = \begin{pmatrix} v_{x0}, v_{x1}, v_{x2}, \ldots, v_{xT} \\ v_{y0}, v_{y1}, v_{y2}, \ldots, v_{yT} \\ v_{z0}, v_{z1}, v_{z2}, \ldots, v_{zT} \end{pmatrix}
\]

\( T \) means transposition, and x-y space frequency \( v_{ij} \) is 2-dimensional DCT values of 64x64 pixel rectangle which is positioned at \( i \times 64 \) of x-coordinate of each frame and \( j \times 64 \) of y-coordinate of each frame (which means the left-top coordinates of the rectangle is \((i \times 64, j \times 64)\)). The number of 2-dimensional DCT values of 64x64 pixel rectangle is 64x64=256. 256 is quite large, so we use three means of low-frequency, middle-frequency, and high-frequency. The mean value of low-frequency is the mean of the values at \((0,1)\), \((1,0)\), and \((1,1)\). The mean value of middle-frequency is the mean of the values at \((32, i)\) and \((j, 32)\) \((i=0, 1, 2, \ldots, 63; j=0, 1, 2, \ldots, 63)\). The mean value of high-frequency is the mean of the values at \((63, i)\) and \((j, 63)\) \((i=0, 1, 2, \ldots, 63; j=0, 1, 2, \ldots, 63)\). The x-t space frequency \( v_{ijt} \) is three means of low-frequency, middle-frequency and high-frequency as the same manner mentioned above of 2-dimensional DCT values of 64x64 rectangle at the position of \(((i + 1) \times 80, j \times 64)\). The x-t space frequency is calculated from three x-t planes pictured on Fig.1.

![Fig.1 First x-t planes of three y-coordinates.](image-url)
In this figure, \( y_0, y_1, y_2 \) are the positions on the \( y \) coordinates, and we use three combinations of \( y_0, y_1, y_2 \). One is \( y_0 = \frac{1}{2} \times y_{\text{width}}, \ y_1 = \frac{3}{4} \times y_{\text{width}}, \ y_2 = \frac{7}{8} \times y_{\text{width}} \), where \( y_{\text{width}} \) is the vertical length of the image plane. Second is \( y_0 = \frac{1}{6} \times y_{\text{width}}, \ y_1 = \frac{5}{12} \times y_{\text{width}}, \ y_2 = \frac{11}{12} \times y_{\text{width}} \). Third is \( y_0 = \frac{1}{6} \times y_{\text{width}}, \ y_1 = \frac{7}{12} \times y_{\text{width}}, \ y_2 = \frac{5}{6} \times y_{\text{width}} \).

Fig.2 shows the figure of \( x-t \) planes of second selection of \( y \)-coordinates, and Fig.3 shows one of third selection. These combinations are selected, because the positions of planes of which time space frequency is calculated are the key values to examine the similarities of time-sequential frames. For example, time frequency of the first combination of three planes from the typical shooting scene is very similar to time frequency of the second combination of three planes from the other similar scene. So we use all combination of these positions to examine the similarity of scenes. In Eq.3, \( v_{ijxt} \) \((i=0,1,2), \ (j=0,1,2,3,4)\) means \( j \)-th DCT values of \( x-t \) plane of \( y_r \)-coordinate.

The \( y-t \) space frequency \( v_{ijyt} \) is three means of low-frequency, middle-frequency and high-frequency of 2-dimensional DCT values of \( 64\times64 \) rectangle at the position of \( (j\times64,(i+1)\times60)-(j\times64+1,(i+1)\times60) \). Also the \( y-t \) space frequency is calculated from three combinations of \( y-t \) planes in the same manner of \( x-t \) planes. Fig.4 shows the example of \( y-t \) planes.

We use time-sequential frames of \( 320\times240 \) pixels captured from broadcast soccer videos. After all, the number of elements of each feature vector is 141, and we extract nine set (three combinations of \( x-t \) planes \( \times \) three combinations of \( y-t \) planes) of feature vectors from each frame.

All above descriptions are of full-sized (320x240) images. In our proposed method, we use three sized images (full, half, and quarter sizes) of image pyramid in our experiments.

4 Eigenspace of Feature Vectors

For selecting personal favorite scenes, we apply the eigenspace of collected feature vectors from learning samples of frames captured from broadcast soccer video. All feature vectors are regularized before following calculation. \( m \) is the mean of all feature vectors. You can make the eigenspace of all learning feature vectors as following manner.

\[
\Sigma = E\{(v - m)(v - m)^T\} \tag{5}
\]

If \( 1 \leq i \leq d = 141 \), then the solution of the following eigen problem (6) can make the subspace of \( k \) eigenvectors corresponded by \( k \) largest eigenvalues (\( k < d \)).

\[
\Sigma u_j = \lambda_j u_j \tag{6}
\]

Each feature vector can be projected to the position in the subspace above as the vector \( y \).

\[
y = (u_0, u_1, \cdots, u_{k-1})^T v \tag{7}
\]

Each user can select the favorite scenes as the group of frames that the user points out by hand. For example, user A has selected aggressive shooting scene as the frame number \( N1 \) to \( N2 \), corresponding feature vectors of the frame number \( N1 \) to \( N2 \) are projected in the subspace calculated above as the vectors from \( y_{N1} \) to \( y_{N2} \). These vectors \( y_{N1} \) to \( y_{N2} \) are the representation of user A’s favorite scenes in the subspace.

After these learning process, you can select the user A’s favorite scenes as the collection of frames which subspace vectors \( y_{A’s favorite} \) satisf
fied by the following equation.

\[ \text{distance}(y_1, y_2) \leq \text{dis}_{\text{threshold}} \]

for any A = N1 to N2, where distance(y1, y2) is the distance between y1 and y2. In this paper, we use the square root of inner product of subtractions of two vectors. dis\text{threshold} is the threshold value of the distance.

5 Experiments and Discussions

We applied the proposal method to a broadcast soccer video which resolution is 320x240, the frame rate is 30/sec. And we use the image pyramid of three resized images, 320x240, 160x120, 80x60. The subspace has been made by the feature vectors of learning video which includes five shooting scene and one goal scene. The total length of video is about 20 minutes = 36000 frames. The total number of shots is 163. And 20501 frames of them are extracted as ‘Playing Field shots’ which are used as input images. These 20501 frames of ‘Playing Field shots’ are divided into 59 shots. User A has selected 106 frames (one shot) as the example of his favorite scene (these frames are in the shooting and goal scene.) Dimension of subspaces k is 40 in full-sized images, 20 in half-sized images, and 10 in quarter-sized images. The distance threshold dis_{\text{threshold}} is 0.3 in full-sized images, 0.5 in half-sized and quarter-sized images. num1_{\text{threshold}} is 4, and num2_{\text{threshold}} is 30% in our experiments. All these threshold values are decided experimentally. According to our experimental results of selecting A’s favorite scenes, selected shots as user A’s favorite scene are six shots. These six shots include similar shooting shots (three shots), similar but non-shooting shots (three shots). In all 59 shots of ‘Playing Field shots’ as input images, there are seven shooting scenes (one of them is a goal scene which is selected by user A as the favorite scene). Three shots of seven include goal scenes, and other two of seven are quite aggressive shooting scenes similar to the goal scene, but the rest two shooting scenes are not similar to the goal scene (one is the scene of the long kick shoot, and the other is the scene that the length is very short.) Then the ground truth of user A’s favorite scenes includes first three shooting scenes and second two quite aggressive shooting scenes. Comparing this ground truth with the results of our method for selecting A’s favorite scenes, five selected shots are true positives and one rest shot is false positive. And there is no true negative.

Recall and precision of A’s experiments are as follows. recall = 5/5 = 100%, precision = 5/6 = 83.3%. Another sample of viewers B has selected one passing ball scene as his favorite scene. In this experiment, recall is also 100%, but precision becomes approximately 70%. In our experiments, we used nine combinations of time-space images for calculating time-space frequency. Considering the results of our experiments, these nine combinations are quite effective for selecting viewers’ favorite scenes. For example, selections of y-coordinates of x-t planes in Fig.1 to Fig.3 are considered as motion characters of middle portion of images, upper portion of images, and lower portion of images in vertical directions respectively. And each scene can be characterized by the combination of these motion characters for selecting personal favorite scenes.

6 Conclusions

In this paper, we present a novel framework to selecting personal favorite scenes using principal component analysis. This time, we implemented this framework only with image space DCT and time space DCT, and estimate moderately for the training data only from these two of information. If color histograms or audio power spectrum are extracted, our framework can be easily extended by adding them to the feature vectors, and we expect more effective results of scene selection of personal favorite scenes. Our future works are to introduce color histograms and audio information, and to prepare more training and testing data.

References

