Improving Performance using both of Correlation and Absolute Difference on Similar Play Estimation

Kyota Aoki and Ryo Aita
Graduate school of Engineering, Utsunomiya University, Utsunomiya, Tochigi, JAPAN

Abstract - Plays in sports are described as the motions of players. This paper proposes the similar play retrieving method based on the motion compensation vector in MPEG sports videos. In MPEG videos, there are motion compensation vectors. Using the motion compensation vectors, we don’t need to estimate the motion vectors between adjacent frames. This work uses the 1D degenerated descriptions of each motion image between 2 adjacent frames. Connecting the 1D degenerated descriptions on time direction, we have the space-time image. This space-time image describes a sequence of frames as a 2-dimensional image. Using this space-time image, this work shows the performance using both of the correlation and the absolute difference to retrieve a small number of plays in a huge number of frames based on a single template play. Our experiment records 0.94 as recall, 0.800 as precision and 0.865 as F-measure in 138 plays in 132503 frames.

Keywords: play estimation; motion compensation vector; MPEG video; Absolute difference; correlation

1 Introduction

There are many videos about sports. There is a large need for content-based video retrievals. The amount of videos is huge, so we need an automatic indexing method [1][2][3][4][5]. We proposed the method retrieves shots, including a similar motion based on the similarity of the motion with a sample part of videos [6][7][8][9].

In former works, we used the correlation of motions and the correlation of textures [10]. We have a good performance using both of the correlation of motions and the correlation of textures.

This paper proposes the method to retrieve the plays using only motion compensation vectors in MPEG videos. Using multiple features, the performance increases. Many works try to index sport videos using the motions in the videos. Many works try to understand the progress of games with tracking the players. Many of the works use the motion vectors in MPEG videos. They succeed to find camera works. They are zoom-in, zoom-out, pan and etc. However, no work retrieves a play of a single player from only motions directly. Of course, camera works have an important role in understanding videos. Sound also has some roles in understanding videos. Many works use camera works and sound for understanding sport videos. Those feature-combining methods have some successes about retrieving home-runs and other plays. However, those works did not success to retrieve plays from only motions. Recently, many works focused on retrieval combining many features and their relations [11][12].

In this paper, we tries to propose the method that retrieves similar plays only from motions. With the help of texture, the performance of similar play retrieval increases. We showed that the combination both of motions and textures shows better performance about similar play retrieval [13]. However, in many games, there are changes of fields and spectators. The colors of playing fields changes with the change of seasons. The motion based method can works with textures, sound, and camera works. However, this paper proposes the method to estimate similar plays only from motions in motion compensation vectors in MPEG videos. There are many motion estimation methods [14][15]. However, they need large computations. The method gets the motions from motion compensation vectors in MPEG videos, and makes the 1-dimensional projections from the X motion Y motion. The motion compensation vector exists at each 16x16 pixels square. It is very sparse description of motions that the motion description made from the motion compensation vector. The 1-dimensional projection represents the motions between a pair of adjacent frames as a 1-dimensional color strip. The method connects the strips in the temporal direction and gets an image that has 1 space dimension and 1 time dimension. The resulting image has the 1-dimensional space axis and the 1-dimensional time axis as the temporal slice [16][17][18]. Our method carries information about all pixels, but the temporal slice method does only about the cross-sections. We call this image as a space-time image. Using the images, the method retrieves parts of videos as fast as image retrievals do. We can use many features defined on images.

The proposed method uses a single template space-time image, and retrieves similar plays as the play described in the template space-time image. A long space-time image template is good for retrieving the same play in the template. However, in a similar play retrieval, the long template of a space-time image is not robust about the change of the duration of a play. A short space-time image template is robust about the change of the duration in a similar play retrieval. However, a short space-time image template is weak in the discrimination power.

This paper use both of correlation and absolute difference in similar play retrieval on only motions. First, we show the over-all structure of the proposed method. Then, space-time
image is discussed. Next, we discuss the similarity measures based on correlation and absolute difference. Then, we shows the experiments on a video of a real baseball game on a Japanese TV broadcast. And last, we conclude this work.

2 Structure of the proposed method

The proposed play retrieving method is the composition of a correlation and an absolute difference of motion space-time images [6][7][8][9][10]. We describe the space-time image as ST-image in the followings. Fig. 1 shows the over-all structure of the proposed method. In fig. 1, the left side is the correlation based retrieval and the other side is the absolute difference based retrieval. In MPEG video, each 16x16 pixel block has the motion compensation vector. Using the MPEG motion compensation vector as the motion description, the size of the description is 1/64 of the original frames. In other word, the static texture description is 64 times larger than the motion descriptions in size. The static texture description is 96 times larger than the motion description in total amount with the consideration of the vector size.

The color description fits for describing the static scenes. The motion description does for describing the dynamic scenes. A play in a sport is a composition of motions. So, the motion description fits for describing plays in a sport. For retrieving the same play in a video, we can use any similarity measures. However, our goal is the retrieval of similar plays from videos. The speed of a play may change. In the case, the similarity measure must be robust about the absolute amount of motions. Our motion retrieval method uses the simple correlation as the similarity measure. The simple correlation works well in the similar play retrieval in sport videos. Using correlation as the similarity measure, there is no hint for the absolute amount of motions. As a result, there are some error retrievals. In the relative motion space, there is no difference between large motions and small motions.

The absolute difference is a major difference feature. In similar play retrieval, the absolute difference shows poor performance in our preparation experiments. Table I shows the experimental results.

The absolute difference only shows poor performance. However, with other similarity measure the absolute difference can work some role for distinguish the similarity measured with the correlations.

The proposed similar play retrieving method uses both of the correlation and the absolute difference between a template ST-image and the ST-image from videos. We can easily differentiate the large motion difference with the absolute difference. In sport videos, the absolute difference can find the camera motions. Composing the correlation based measure and the absolute difference based measure, we construct the composite similar play retrieval method.

3 1D Degeneration from videos

There are many degeneration methods. Some works use the temporal slices [13][19][20]. The temporal slices are easy to make and represent videos in small representation. The temporal slice is the sequence of the set of selected pixels in frames. There is no information about other pixels. The previous works make the temporal slices from color and textures in videos.

This paper makes 1D degenerated representation using the statistical features of the set of pixels. The main statistical features are mean, mode and median. This paper uses the mean for making 1D degeneration. For treating sports, the motions in videos are important.

3.1 Motion Extraction from MPEG videos and construction of space-time image

We show the process creating the space-time images of motions in an MPEG video in figure 1. From an MPEG video, we get a 2-dimensional image describing motion compensation vectors in a video. In the following, we abbreviate space-time image as ST image. The ST image has a space axis and time axis. The ST image represents the sequence of frames in one 2-dimensional image. The ST image is a very compact description of a video, and it is easy to treat.

3.2 Motion extraction from MPEG videos

First, we must have the motions in an MPEG video. An MPEG video is a sequence of GOP (Group of Pictures). Each GOP is starting from I-frame, and has B-frames and P-frames as shown in figure 2. The motion compensation vector is block-wise. The block size is 16 x 16. There is a 640 x 480 MPEG video. The motion compensation vector image is only 40 x 30 pixels.
In the color frame, the grids show the motion compensation blocks. In the motion compensation vector image, the intensity of red shows the X-direction motion, and the one of green does the Y-direction motion.

MPEG videos have 3 types of frames. They are Intra-coded frame, Predicted frame and Bi-directional predicted frame. Intra-coded frame is called I-frame. Predicted frame is P-frame. Bi-directional coded frame is B-frame. The I-frame is an independent closed coded frame. There is no motion compensation vector. The P-frame has a forward prediction. The B-frame has forward and backward predictions.

There is no motion compensation vector in I-frame. The B-frame before the I-frame has a forward and a backward motion compensation vector. We use the reversed backward motion compensation vector for the motion vector of the I-frame. The P-frame has a same problem. We also use the reversed backward motion compensation vector in the B-frame just before the I-frame.

3.3 Space-time image

We have the motion vector (2-dimensional) on every motion compensation block. The amount of information is 2/16 × 16 × 3 of the original color video. This is very small comparing with the original color frames. The base-ball games can long about 2 hours. This video has 200K frames. If we compare frame by frame, there needs a huge computation. There is a large difficulty to retrieve similar parts of a video.

We can retrieve similar parts of videos using classical representative frame-wise video retrieve method. However, it is difficult to retrieve similar part of videos based on the player’s motions, because the motion leads a change of subsequent frames.

We can use many feature extraction methods to retrieve similar part of videos, but the applicability of the method depends on the features selected to use. The generality of the method may be lost using specified features.

This paper uses the 1-dimensional degeneration for reducing the amount of information without lost generality [3]. Fig. 2 shows the process to create a ST image from motion vector frames. We make 1-dimensional degeneration of each frame as in the top of fig. 2.

\[ I_{1dx}(x) = \frac{\sum_{y \in [0,Y_{max}]} I_{2d}(x,y)}{Y_{max} + 1} \] (1)

\[ I_{1dy}(y) = \frac{\sum_{x \in [0,X_{max}]} I_{2d}(x,y)}{X_{max} + 1} \] (2)

Equation (1) makes 1-dimensional degenerated description of X direction from a 2-dimensional image. Equation (2) makes 1-dimensional degenerated description of Y direction from a 2-dimensional image. In (1) and (2), \( I_{2d}(x,y) \) stands for the intensity at the pixel \((x,y)\). \((X_{max}, Y_{max})\) is the coordinate of the right-upper corner.

The resulting 1-dimensional degenerated description is defined as (3).

\[ I(j) = \begin{cases} j \leq X_{max} & \rightarrow & I_{1dx}(j) \\ j > X_{max} & \rightarrow & I_{1dy}(j - X_{max} - 1) \end{cases} \] (3)

There are 2 directions to make a 1-dimensional degeneration. We use both 2 directions that are X-axis and Y-axis using (1) and (2). In each color plane, we have a 1-dimensional degenerated description. We connect the 2 degenerated descriptions onto X-axis and the transposed projection onto Y-axis as from the second to the third of figure 4. We represent the X-direction motion in red, and Y-direction motion in green. There is no value in blue. Then, we have a 1-dimensional degenerated color strip from the motion compensation vector. In the color strip, red represents the X-direction motion and green does the Y-direction motion. For the convenience, we set 255 in blue when both of X and Y direction motions are 0.

We connect the 1-dimensional color strips describing motion frames on time passing direction as the bottom of figure...
4. Connecting 1-dimensional color strips, we have a color image that has 1 space axis and 1 time axis. In this paper, the image is described as ST (Space-Time) image. In the following experiments, we use the 320 × 240 pixels half size frames. In the MPEG format, each 16 × 16 pixels block holds a motion compensation vector. This leads to reduce the amount of information into 1/256. The resulting motion image is 20 × 15 pixels. The 1D degenerated description is 7/60 of the original 20 × 15 pixels image. As a result, the usage of the ST image of motion compensation vector in MPEG video reduces the amount of information into 0.045% from the original half size video frames. In the ST image used in this paper, X-axis holds the space and Y-axis does the time. There is no reduction of information in time axis.

The similar motion retrieval estimates what kind of motions exists on a place. It is same as the cost of the retrieval on images to retrieve similar part of videos on a ST image.

There are many similar image retrieval methods. They can be applied in ST images describing videos’ motions. This paper uses the correlation between two images. We normalize the resulting correlations for compensating the variance among videos. We have 2 independent correlations between two ST images from each color plane. The blue plane exists for only the convenience for our eyes. We use an X-direction motion and a Y-direction motion in red and green planes.

Fig. 3 shows the ST-images from motions and colors. The right one is the sequence of the original video’s frames. The left one is the ST-image based on the motion compensation vectors. It is enlarged in space-axis. There is no colors representing a uniform. The center one is the ST-image from the colors for comparing. In the ST-image based on colors, we can see the uniform colors.

3.4 Matching between template ST image and retrieved ST image

All ST images have same space direction size. The estimated motion image is X × Y pixels. Then, the size of the space axis of ST images is X + Y pixels. For computing the correlations between the template ST image and any part of retrieved ST image, there is no freedom on space axis. There is only the freedom on time axis. If a template ST image is S × t and a retrieved ST image is S × T, the computation cost of correlations is S × t × T.

In a baseball game, the length t of an interesting play of a video is short. So, the computational cost of correlations is small enough to be able to apply large scale video retrieval. Because of the shortness of the retrieved part, there is no need to compensate the length of the part. There is no very slow pitch or no very fast one. There is no very slow running or no very fast one.

4 Similarity measure with correlation and absolute difference in motion retrieving method

4.1 Similarity measure in motion space-time image based on correlations

We use the mutual correlation as the measure of similarity. We have 2 dimensional correlation vectors. They are X-direction motion, Y-direction motion. If there is a similar motion between the template and the retrieved part of ST image, both of the 2 correlations are large. We use the similarity measure shown in (4).

\[ S_C(I_T, I_V) = \min_{p \in [x, y]} (NCol(I_{T_p}, I_{V_p}) - Th_p) \]  

\[ NCol(I_{T_p}, I_{V_p}) = \frac{\sum_{I_{V_p} \in V} Col(I_{T_p}, I_{V_p}')} {SD_{I_{V_p} \in V}(Col(I_{T_p}, I_{V_p}'))} \]  

In (4), \( S_C(I_T, I_V) \) is the similarity between a ST-image \( I_T \) and \( I_V \). \( I_T \) is a template ST-image. \( I_V \) is a part of ST-image from a video. \( NCol(I_{T_p}, I_{V_p}) \) is the normalized correlation between \( I_{T_p} \) and \( I_{V_p} \) over a video \( V \). The normalized correlation is normalized on the pair of the template ST-image and the ST-image from a video over the video. \( Th_p \) is the threshold. \( p \) is one of \( x \) and \( y \) that represent the X-direction motion and Y-direction motion. This similarity measure is scalar.

The equation (5) is the formal definition of \( NCol \). \( SD \) is a standard derivation over the video \( V \). \( Col \) is the correlation.
4.2 Similarity measure in motion space-time image based on absolute differences

We use the absolute difference as the measure of similarity in the motion description. In the case, we have 2 dimensional absolute differences. They are the absolute differences based on both of X-direction motion and Y-direction motion. If there is a similar motion between the template and the retrieved part of the ST image, both of the 2 absolute differences are small. We use the similarity measure shown in (6).

\[
S_D(I_o, I_1) = \min_{p \in \{x, y\}} (NABD(I_{op}, I_{1p}) - T_h_p)
\]  

(6)

\[
NABD(I_{Tp}, I_{Vp}) = \frac{\sum_{I_{v'p} \in V} ABD(I_{Tp}, I_{v'p})}{|V|} - ABD(I_{Tp}, I_{Vp})
\]  

(7)

In (6), \(NABD(I_{op}, I_{1p})\) is the normalized negation of absolute difference between \(I_{op}\) and \(I_{1p}\). \(I_{op}\) and \(I_{1p}\) are ST images. \(T_h_p\) is the threshold. \(p\) is one of \(x\) and \(y\) that represent the X-direction motion and Y-direction motion. This similarity measure is scalar.

The equation (7) is the formal definition of \(NABD\). \(SD\) is a standard derivation over the video \(V\). \(ABD\) is the absolute difference between 2 ST-images. In (7), the terms in the numerator are placed reverse order from a normal position. This implements the negation.

5 Experiments on baseball games and evaluations

5.1 Baseball game

This paper treats baseball game MPEG videos. In baseball games, players’ uniforms change between half innings. The pitch is the most frequent play in a base-ball game. There is large number of pitches. This paper uses a single play of a pitch as a template. Using this template, the proposed method retrieves large number of pitches using similar motion retrieval.

Motion based similar video retrieval can find many types of plays based on the template. There are a few repeated plays that are not pitches. This paper distinguishes a pitch and other plays.

5.2 Experimental objects

This paper uses a whole base-ball game for experiments. The game is 79 minutes, 132485 frames in a video. In the game, there are right-hand pitchers and a left-hand pitcher. There are 168 pitches. There are 31648 frames that represent the camera work that catches the pitching scenes.

Fig. 4 shows the example of pitches in our experimental video at each 5 frames distance. The center one and right one differ from the left one at the uniforms. The right one differs from the left one at left-hand and right-hand.

5.3 Experiment process

The experimental videos are recorded from Japanese analog TV to DVD. Then, the recorded videos are reduced into 320 × 240 pixels and encoded MPEG1 format. Most plays of pitches are very short. So there is no reduction on time direction. There are 30 frames in 1S. There are all parts including telops, sportscasters and CG overlays. The first step of our experiment is the extraction of motion compensation vectors. The motion compensation vector is at each 16 × 16 pixel blocks. In every motion compensation block, we have a motion compensation vectors. The similar play retrieval in motion frames uses 20 frames of the pitch as in fig. 2 as a template and retrieves the shots including pitches of a video. In Fig. 3, the left one shows the part of the ST-image based on the motions from motion compensation vectors in a MPEG video. And, the center shows the ST-image based on the original colors in frames. The frames of this part are shown in the right.

Fig. 4 Examples of pitching shots.
5.4 Correlation based similarity measure in pitching retrieval

These experiments use a single template image of the length 20 frames. When we say the template that starts 830000th frame, we use 830000th frame to 830200th frame. There are 20 motions. This sequence has 21 frames. Our pre-experiments using some length of template ST-image show that the 20 motion frames template ST-image is best. Here after, we use the template of the length 20. We control the thresholds that make the F-measure as the maximum using Excel goal seek function.

In this experiment, we use 3 template pitches. They are the template starts from 120130th, 123340th and 191360th. The F-measures spans from 0.503 to 0.807. Table I shows the experimental results. We try some templates. The template started from 120130th frame is best. We select these 3 templates that includes the best one.

In number, in the best case, the method finds 131 pitches in 168 pitches. The method retrieves 157 candidate pitches in 132485 possible candidates. This means that the 26 error retrievals in 132317 no candidates. This is 99.98% precision. In this case, there is huge amount of frames. There is huge unbalance between yes samples and no samples. In the case, it is difficult to get a high F-measure.

5.5 Absolute difference based similarity measure in pitching retrieval

In this experiment, we use the same 3 template pitches in the correlation based method. They are the template starts from 120130th, 123340th and 191360th. The F-measures spans from 0.315 to 0.371. The performance is very low. Table II shows the experimental results. In number, the method finds 165 pitches in 168 pitches in the video. The method retrieves 708 error candidates. This shows that the absolute difference based similarity cannot distinguish a pitch from other motions.

5.6 Combination both of correlations and absolute differences

For retrieving the precise pitches in frames, we need to use the correlation based similarity. With combining the correlation based similarity and the absolute difference based similarity, we can improve the performance of the precise pitch retrieval.

Using the correlation based similarity or the absolute difference based similarity, we have no difficulties to find the proper set of thresholds. In the correlation based similarity or the absolute difference based similarity, there are 2 thresholds that work in each X and Y direction motions.

To combine the correlation based similarity and the absolute difference based similarity, we use logical conjunction. We can use logical disjunction. However, minimum is extension of logical conjunction. We select logical conjunction in this paper. We have 4 thresholds in the combination. It is difficult to optimize all thresholds at once with short processing time. We divided the optimization of thresholds into 2 steps. They are the correlation based similarity threshold optimization and the absolute difference based one.

First, we optimize the thresholds in the absolute difference based similarity. Then, we do ones in the correlation based similarity. In the absolute difference based experiment, the recalls are high enough. This leads the decision. Of cause, there is no difference between the correlation based first and the absolute difference based first.

Table III shows the result of the experiments. The templates are same in former experiments. The resulting F-measures spans from 0.745 to 0.865. In every template, the performance increases much. Especially, in the case of the template starting 191360th, there is 48% increase in F-measure.

6 Conclusions

This paper discusses about the retrieval of similar plays in sport MPEG videos using similar motion retrieval based on both of correlation and absolute difference. For recognizing sport videos, the motions represent important meanings. In the cases, there must be similar video retrieval methods based on the motions described in the videos. The proposed similar play retrieval method is the combination of correlation and absolute difference motion based on only motion compensation vectors in MPEG videos.

<table>
<thead>
<tr>
<th>Template</th>
<th>Recall</th>
<th>Precision</th>
<th>F-measure</th>
</tr>
</thead>
<tbody>
<tr>
<td>120130</td>
<td>78.0%</td>
<td>83.7%</td>
<td>0.807</td>
</tr>
<tr>
<td>123340</td>
<td>70.8%</td>
<td>77.8%</td>
<td>0.742</td>
</tr>
<tr>
<td>191360</td>
<td>58.9%</td>
<td>43.9%</td>
<td>0.503</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Template</th>
<th>Recall</th>
<th>Precision</th>
<th>F-measure</th>
</tr>
</thead>
<tbody>
<tr>
<td>120130</td>
<td>98.2%</td>
<td>18.9%</td>
<td>0.371</td>
</tr>
<tr>
<td>123340</td>
<td>100.0%</td>
<td>18.7%</td>
<td>0.315</td>
</tr>
<tr>
<td>191360</td>
<td>100.0%</td>
<td>18.8%</td>
<td>0.316</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Template</th>
<th>Recall</th>
<th>Precision</th>
<th>F-measure</th>
</tr>
</thead>
<tbody>
<tr>
<td>120130</td>
<td>94.0%</td>
<td>80.0%</td>
<td>0.865</td>
</tr>
<tr>
<td>123340</td>
<td>83.9%</td>
<td>87.9%</td>
<td>0.859</td>
</tr>
<tr>
<td>191360</td>
<td>84.5%</td>
<td>66.6%</td>
<td>0.745</td>
</tr>
</tbody>
</table>
The experiment shows that the similar play retrieval works well using both of correlation and absolute difference on ST-images made from motion compensation vectors in MPEG videos. Classical works using MPEG motion compensation vector only uses global-scale motions. However, the proposed method utilizes local motions. The proposed combination of correlation based similarity measure and absolute difference based similarity measure works well in our experiments. Using both similarity measures, the proposed method gets some more performance than a single correlation based similarity measure base on motion based ST-images.

REFERENCES


