Spatio-Temporal Pyramid Matching for Sports Videos

Jaesik Choi
University of Illinois at Urbana-Champaign
Urbana, IL
jaesik@cs.uiuc.edu

Won J. Jeon
University of Illinois at Urbana-Champaign
Urbana, IL
wonjeon@illinois.edu

Sang-Chul Lee
Inha University
Incheon, Korea
sclee@inha.ac.kr

ABSTRACT

In this paper, we address the problem of querying video shots based on content-based matching. Our proposed system automatically partitions a video stream into video shots that maintain continuous movements of objects. Finding video shots of the same category is not an easy task because objects in a video shot change their locations over time. Our spatio-temporal pyramid matching (STPM) is the modified spatial pyramid matching (SPM) [12], which considers temporal information in conjunction with spatial locations to match objects in video shots. In addition, we model the mathematical condition in which temporal information contributes to match video shots. In order to improve the matching performance, dynamic features including movements of objects are considered in addition to static features such as edges of objects. In our experiments, several methods based on different feature sets and matching methods are compared, and our spatio-temporal pyramid matching performed better than existing methods in video matching for sports videos.

Categories and Subject Descriptors

I.4.7 [**IMAGE PROCESSING AND COMPUTER VISION**]: Feature Measurement; I.2.10 [**ARTIFICIAL INTELLIGENCE**]: Vision and Scene Understanding—*Motion*

General Terms

Algorithms

Keywords

content-based video retrieval, spatio-temporal, pyramid matching, video partitioning

MIR'08, October 30–31, 2008, Vancouver, British Columbia, Canada.

1. INTRODUCTION

Content-based video retrieval systems have different characteristics from content-based image retrieval systems. As Marchand-Maillet[17] and other researchers pointed out earlier, the research on video retrieval systems mainly targets three challenges - (1) how to use temporal information (such as motion), (2) how to query the system with ease, and (3) how to organize the information of videos.

- **Temporal analysis**: Temporal analysis on different levels of videos such as scene, shot, video[1], etc. focuses on the temporal features such as motion and audio, and is followed by indexing such as key frame detection using these features. Motion analysis characterizes the statistics of global (major object or camera motion) and object motion.
- **Queries to video**: Conventional text queries to content-based image or video documents are difficult to solve, so a query-by-example is popular approach in content-based image and video retrieval systems. Depending on interesting features of visual documents, there have been different approaches such as visual query, motion query, textual query, and the combination of these queries.
- **Organization of video information**: Metadata information from the temporal analysis of videos is stored along with the videos in database in order for querying a new video.

From users' point of view, querying video retrieval system has been quite primitive and limited. There are two main reasons why these types of simple querying systems have been used despite of its shortcomings. First, introduction of temporal information on videos adds more complexity to dimensionality of data, so queries could be more complex than typical text-based ones. On the other hands, representing these queries generated by simple sketch tools are so primitive or generic compared with text-represented queries, they would lead either wrong or diverse query results. In addition, more complex querying system (such as dynamical construction of hierarchical structures on targeting videos) requires more elaboration on queries by users, which could be more error-prone. Second, it has been assumed that users do not have sample videos at hand for query, so additional

[1]In this paper, a *scene* is an image frame, a *shot* is a set of image frames which has continuous movement of objects in it, and a *video* is a set of shots.

querying tools are required. However, this assumption is no longer valid because mobile devices such as digital cameras, PDAs, and cell-phones with camera and solid-state memory enable instant image and video recording which can be used for video query.

Our content-based video query system takes a sample video as a query and searches the collection of videos typically stored in multimedia portal service such as YouTube[29], Google Video[8], Yahoo! Video[28], and MSN Video[18]. It suggests similar video clips from the database with relevance evaluation. The whole system works in two phases - (1) *video partitioning* for a new video entry into the database system and (2) *video matching* process for a new query video. When a video is stored in the database, it is partitioned into multiple shots by our shot boundary detection based on feature analysis and classification of the features. The partitioned video shots are stored along with metadata information in the database. In query process, a new video is analyzed and matched to the stored videos, and the relevant scores are calculated by using our spatio-temporal pyramid matching method.

The rest of this paper is organized as follows. In section 2, related work on image and video matching is discussed. Our spatio-temporal pyramid matching is presented in section 3. We provide mathematical discussions on spatio-temporal pyramid matching in section 4. The experimental settings and results are presented in section 5 followed by our conclusion.

2. RELATED WORK

The challenges and characteristics of content-based image and video query systems are well discussed in [17][13]. Ulges *et al.*[24] and Ando *et al.*[2] discussed video tagging and scene recognition problems, which have similar goals to ours but takes different approaches. The techniques to summarize features in videos using hidden Markov model have been used in [3][14][2]. Compared with using hidden Markov model, our modified pyramid matching scheme has the simpler representation of features in time domain and therefore is faster to calculate the score of relevance feedback for video query.

Recently content-based video retrieval for sports video has been widely discussed. The work in [23] and [31] focused on the framework and personalization of generic sports videos whereas the other works target particular sports such as baseball[19][7], soccer[27], basketball[26], etc. Our system has a general framework which is applicable to any type of videos for matching and querying.

Different techniques in finding similarity of subsets of video streams have been discussed in [22][1][30][21][10]. Pyramid matching[9][12][25] is believed as one of the best matching algorithms for image retrieval and recognition. Our spatial-temporal pyramid matching algorithm extends pyramid matching to time domain for efficient video matching and query. In addition, we provide the mathematical condition for superiority using time domain.

Boosting algorithm[6] is a general framework for performing supervised learning. Recently the work in [20] used it for keyframe-based video retrieval, whereas we use it for weighing different feature sets to determine matching scores.

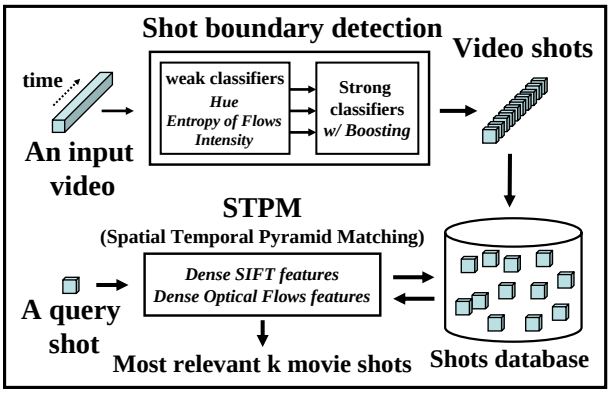

Figure 1: A diagram of our system

3. SYSTEM DESIGN

Our system is composed of two parts as show in Figure 1 - (1) automatic detection of shot boundaries and (2) similarity matching of video shots. Given a video file, our system divides the video into video shots based on automatically detected boundaries. The similarity of video shots is measured by our spatio-temporal pyramid matching and it is used as the rank of video matching in video query.

In this section, we present our spatio-temporal pyramid matching for video matching problems, followed by the details of our system design including weight assignment on features, shot boundary detection, and shot similarity matching.

3.1 Spatio-Temporal Pyramid Matching

Spatial pyramid matching [9] has been successfully applied to recognize images of natural scene. In pyramid matching[12], the pyramid match kernel is built by constructing a pyramid with L levels and finding the number of matches using histogram intersection function [9], followed by weighing $\frac{1}{2^{l-1}}$ in each level l.

The resulting kernel κ^L is defined as:

$$\kappa^L(X,Y) = \Gamma^L + \sum_{l=1}^{L-1} \frac{\Gamma^l - \Gamma^{l+1}}{2^{L-l}} = \sum_{l=1}^{L} \frac{1}{2^{L-l+1}}\Gamma^l + \frac{1}{2^L}\Gamma^0 \quad (1)$$

where Γ_l is the number of matched features in the level l between two images (X and Y).

Different from images, videos have additional information on time domain in addition to spatial domains. Figure 2 shows the hierarchical structure in both spatial and temporal domains in video shots for our Spatio-temporal pyramid matching. Spatio-temporal pyramid matching partitions a video shot into 3D grids in spatial and temporal space and calculates the weighted sum of number of matches in each level. A sequence of grids at level $0, ..., L$, is constructed such that the unit grid at level l where $l \in [0, L]$ has 2^l cells in each dimension. Let ${H^l}_X$ and ${H^l}_Y$ denote the histograms of two images X and Y at level l, and ${H^l}_X(i)$ and ${H^l}_Y(i)$ are the numbers of points in X and Y that fall into the i_{th} cell of the grid. Then the number of matches at level l is given by the histogram intersection function which is the sum of the minimum value of two histograms in each bin.

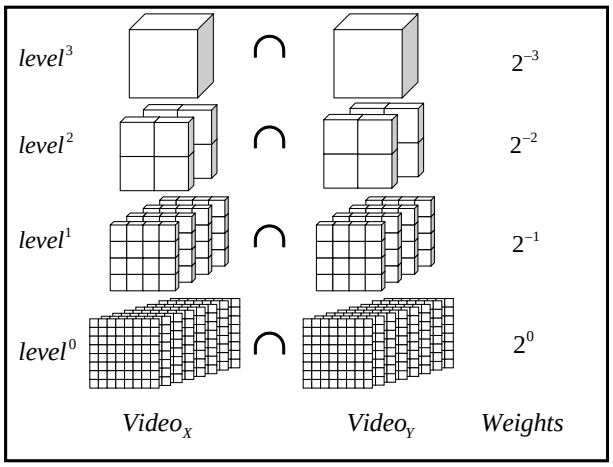

Figure 2: Hierarchical structure in Spatio-Temporal pyramid matching

3.2 Weight Assignment on Features

The performance of video matching is mainly determined by the selection of features and the weight assignment on those features. Two channels of features using SIFT and optical flows are used to match two video shots. In addition, we assign weights of channels of features by using the linear sum of two features[2]. The matching of cubes from two images X and Y are conducted as shown in Figure 3. In this example, the similarity between two video shots X and Y (STPM) is calculated as:

$$STPM(X,Y) = w_{SIFT} * STPM_{SIFT}(X,Y) + w_{OP} * STPM_{OP}(X,Y) \quad (2)$$

3.3 Shot Boundary Detection

To detect shot boundaries in videos, we use our boosting algorithm that combines weak classifiers[3]. Boasting is a well known method to find appropriate weights for each weak classifier. In our implementation, three conventional features including hue channel, entropy of optical flows, and grayscale intensity are used, based on the characteristics of sports videos. However, any additional feature can be pluggable to our framework.

3.4 Shot Similarity Matching

Considering the characteristics of sports videos, we use *SIFT*[15] features and *Optical Flows*[16] to measure the similarity of given two video shots. In order to construct the pyramidal structure in temporal domain, we extract m frames (i.e., 2^l) from each video shot. Thus, the video shots provide a cube of size 2^l by 2^l by 2^l in spatial and temporal space. Each cell in the cube includes the histograms of features. In our implantation, two representative features, dense SIFT and dense Optical Flow were used.

[2]Note that a boosting algorithm could be used if several channels of features beyond SIFT and optical flows are considered.

[3]Weak classifiers are used in this paper to refer threshold-based linear classifiers. They are components of the strong classifier used to determine shot boundaries.

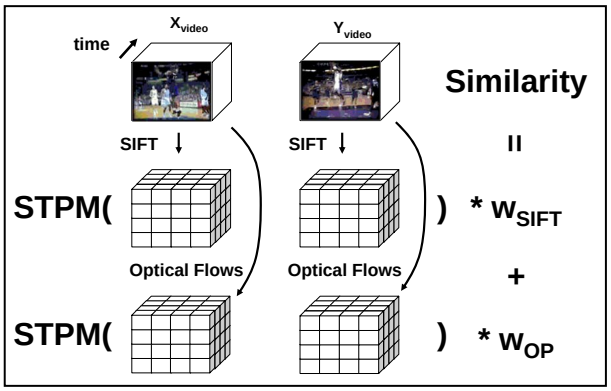

Figure 3: Video matching scheme for spatio-temporal pyramid matching

- **Dense SIFT**: The dense SIFT features are a set of SIFT features which are extracted uniformly in each spatial location. Normally, SIFT features are extracted at the salient locations. However, dense sift features are proven as quite effective in image similarity matching [4, 12]. The SIFT features are clustered into arbitrary number of groups with k-means clustering (e.g., $k = 200$).

- **Dense Optical Flows**: In each location, an optical flow is extracted even though the location has no salient feature. The optical flows are calculated between one of m frames and the next frames in the original video shots. The flows are also clustered into arbitrary number of groups with k-means clustering (e.g., $k = 60$).

4. DISCUSSIONS

In this section, we discuss the mathematical condition in which the temporal information contributes to the matching performance.

Suppose that we extract a set of key images of from two video shots. Given these two sets, the matching score using spatio-temporal pyramid matching (SPTM) is calculated and compared with that of spatial pyramid matching (SPM). In case of SPM, the score is calculated using the conventional weighted sum of each key image. It is straightforward to prove that the matching score of STPM is greater than or equal to the matching score of SPM. However, STPM also provides the higher matching score between two shots in different categories. Intuitively, the gain of STPM compared to SPM in the same category needs to be bigger than the noisy gain[4] in different categories. Otherwise, the performance of STPM could be worse than that of SPM.

4.1 The Gain of Noisy Matching

The expected noisy matching of SPM and STPM in each level as Γ^l_{SPM} and Γ^l_{STPM} are calculated from the equation (1) and shown in Figure 4. Note that we assume that an image has 32 by 32 grid cells (therefore, 1024 features per image), and features (such as SIFT) are clustered into 200 groups (k-means clustering of features for example). In the setting, a grid cell has a feature. In a pyramid level l,

[4]A *noisy* matching score is a matching score between two video shots in different category.

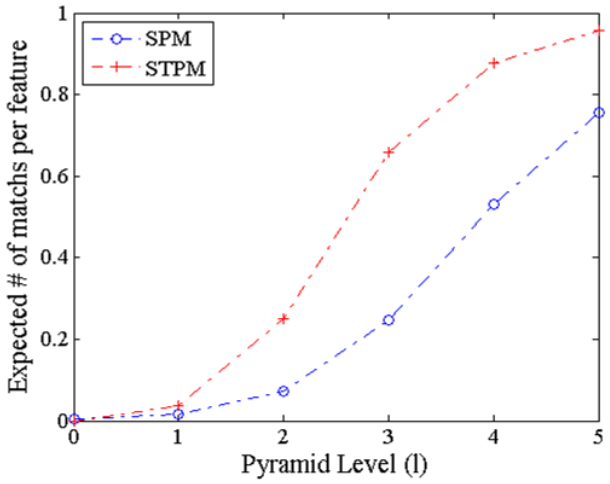

Figure 4: Expected noisy matches per feature

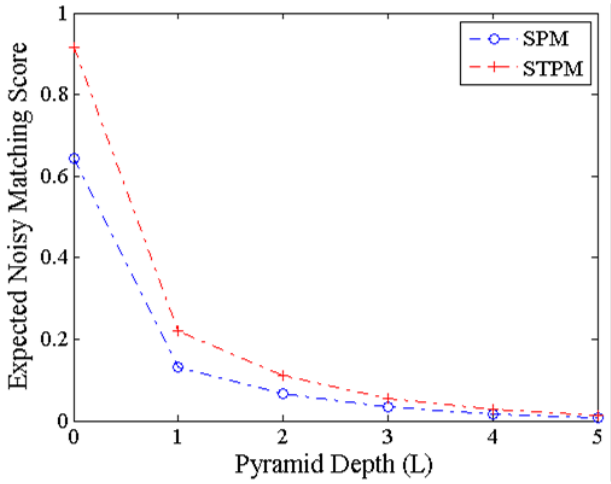

Figure 5: Expected noisy matching score

features of a grid with the size of 2^l by 2^l are aggregated to be matched with the histogram intersection. In case of videos, features of a cube with the size of 2^l by 2^l by 2^l are aggregated. Thus, it could be the bag-of-features in the highest pyramid level (level 5 of Figure 4) which aggregates the whole image (or the video). It is natural to see the two observations: (1) Γ^l_{SPM} and Γ^l_{STPM} (expected number of matched features per feature) increase along the higher pyramid level as shown in Figure 4, and (2) the expected score of $STPM$ is higher than the average score of SPM in each level. The two observations is caused by the same reason. If the aggregated grid is bigger, the histogram intersection per feature also becomes higher.

In case of SPM and STPM, the expected noisy matching score is as follows.

$$\tau^L_{SPM} = \frac{1}{2^L}\Gamma^0_{SPM} + \sum_{l=1}^{L} \frac{1}{2^{L-l+1}}\Gamma^l_{SPM} \quad (3)$$

$$\tau^L_{STPM} = \frac{1}{2^L}\Gamma^0_{STPM} + \sum_{l=1}^{L} \frac{1}{2^{L-l+1}}\Gamma^l_{STPM} \quad (4)$$

As shown in the result in Figure 5, the matching score of STPM is higher than that of SPM in each pyramid level (L). In the minimum pyramid depth, 0, the pyramid has only one level, so that the matching score is same with the bag-of-features assumption. In the maximum pyramid depth, 5, the matching score is dominated by the score of Γ^5 which is more sensitive to spatial and temporal locations of features. The result is caused by Γ^l_{SPM} and Γ^l_{STPM} in Figure 4.

4.2 Conditions for Superiority of STPM

Suppose we have $\kappa^L_{SPM}(A_i, B_i)$ and $\kappa^L_{STPM}(A, B)$, the sound matching scores of SPM and STPM of two videos A and B in the same category. A_i and B_i denote the i_{th} image in the video shots. Then, the following equation is normally satisfied because the size of a cube in video is bigger than the size of a square in image.

$$\kappa^L_{STPM}(A, B) - \frac{\sum_{i \in I} \kappa^L_{SPM}(A_i, B_i)}{|I|} > 0 \quad (5)$$

Here, we refer the value of equation (5) as the gain of $STPM$ compared to SPM. In order to guarantee the better performance of $STPM$ against SPM, the gain should be bigger that the gains of noisy matchings ($\tau^L_{STPM} - \tau^L_{SPM}$) as follows.

$$\kappa^L_{STPM}(A, B) - \frac{\sum_{i \in I} \kappa^L_{SPM}(A_i, B_i)}{|I|} > \tau^L_{STPM} - \tau^L_{SPM} \quad (6)$$

That is, the gain of scores in the same category (left hand side of equation (6)) needs to be greater than the gain of score in different categories (right hand side of equation (6)). In another words, SPM should have room to be improved by the gain as follows, because the maximum score of $\kappa^L_{STPM}(A, B)$ is 1.

$$1 - (\tau^L_{STPM} - \tau^L_{SPM}) > \frac{\sum_{i \in I} \kappa^L_{SPM}(A_i, B_i)}{|I|} \quad (7)$$

5. EXPERIMENTS

In this section, we describe our experimental settings and results. The performance of video matching with our spatio-temporal pyramid matching is measured in two parameters - (1) the quality of video retrieval and (2) the quality of binary decision.

5.1 Experimental Settings

Several sports highlight videos are downloaded from Youtube. The videos mainly contain different sport categories including basketball, football, baseball, and so on. Among these videos, we chose highlights of basketball dunk shots, basketball field goals, and running actions at collage football. The highlights are constructed from hundreds of different sports videos, thus our experimental data is not biased for some specific videos. The videos are divided into several video shots using our algorithm for shot boundary detection which is described in section 3.3.

Next, the evaluation data set which is composed of 200 video shots was prepared with manual labeling process. In the labeling process, the video shots were categorized based on directions of sporting activities (e.g., running left or right) and camera movements (e.g., focusing up or down). As a result, the number of video shots in each category span from 2 to 36.[5]

[5]The details of labels with the data set are available at

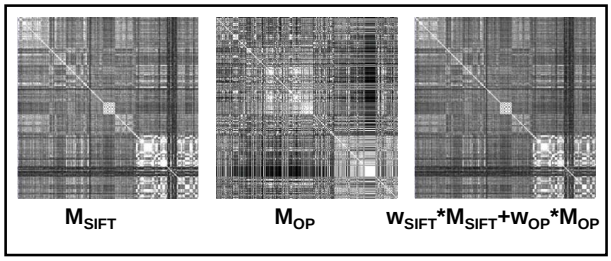

Figure 6: Affinity matrix among video shots

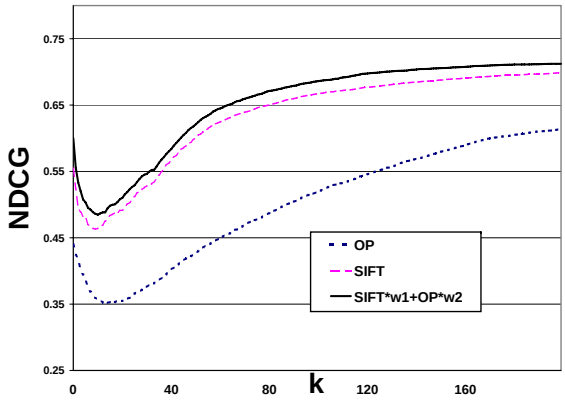

Figure 7: NDCG@k performance for video shot retrieval task

5.2 Experimental Results

5.2.1 *Affinity Matrix*

Given 200 evaluation shots, we calculated the similarity scores among the video shots. We show the result using an affinity matrix [5] as represented in Figure 6. We measure the distance of video shots with three different kernels of STPM (SIFT, optical flow, and the combination of two). X and Y axes of each matrix represent indices of the video shots in our experiments. The white (or black) pixels in each matrix indicates high (or low) similarities between two video shots. Note that the gray scale of each pixel is normalized by the histogram equalization, which results the gaussian distribution ($N(0.5, 0.2^2)$) of the matching scores. In each video shot, the indices are clustered according to the labeled categories of the video shots. It clearly shows that SIFT and optical flows complement each other in some cases.

5.2.2 *Quality of Video Retrieval*

The quality of retrieved videos is measured by NDCG@k (Normalized Discounted Cumulative Gain at top k ranks)[11]. NDCG@k is an evaluation metric in information retrieval for measuring the accuracy of search results. The formula of NDCG is given as follows.

$$NDCG@k = \frac{1}{Z}\sum_{p=1}^{k}\frac{2^{s(p)}-1}{log(1+p)} \quad . \tag{8}$$

http://reason.cs.uiuc.edu/jaesik/viki/supplementary/.

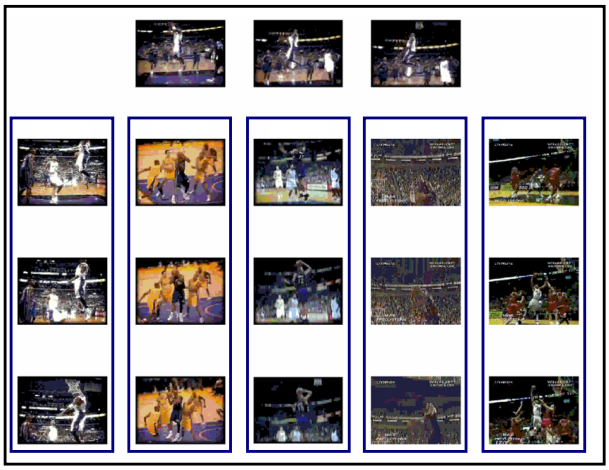

Figure 8: Example of video query and results

In the equation, NDCG@k is the sum of awards from position $p = 1$ to $p = k$ in the ranking results. $s(p)$ is the function that represents rewards given to the video shots at position p. In our experiment, $s(p)$ is the indicator of pth shot whether the shot is in the same category with the query (i.e., $s(p) = 1$ or $s(p) = 0$). The term Z is the normalization factor that makes the perfect result have NDCG@k value of 1.

Figure 7 shows the measurement of three different methods for video retrievals in our spatio-temporal pyramid matching - (1) using only SIFT (*SIFT*), (2) using only optical flows (*OP*), and (3) using combination of weighted SIFT and optical flows (*SIFT*w1+OP*w2*).

Figure 8 shows an example video query and results. For the query, three dunkshot images shown in the top row were used. The five sets of three images on the bottom were shown as querying results in this example. They are sorted in terms of matching ranks of our spatial-temporal pyramid matching from left to right.

5.2.3 *Quality of Binary Decision*

For evaluation purposes, the video query problem is translated into binary decision problem. Given a query shot Q in a particular category, we choose a shot A in the category of the query and a shot B outside the category. Thus, the similarity score, $STPM(Q, A)$ between Q and A should be greater than $STPM(Q, B)$ between Q and B. If $STPM(Q, A)$ is greater than $STPM(Q, B)$, the binary classifier with the matching schema is regarded *correct*. If $STPM(Q, A)$ is less than $STPM(Q, B)$, the binary classifier is regarded *incorrect*. Otherwise, it is regarded *unclassified*.

Our spatio-temporal pyramid matching was compared with two other schemes. As a baseline, we used a single key frame extracted from the center to match the shots whose precision was 75.7%. Then, multiple key frames were used from each video shot. The precision of this method was 84.1%. Finally, our spatio-temporal pyramid matching schema achieved the best matching accuracy of 85.3%.

Figure 9 shows the result of another experiment in quality of binary decisions. Three different schemes were compared in this scenario, (1) using only key frames, (2) matching with only spatial pyramid matching (SPM), and (3) matching

Methods		Correct	Incorrect	Unclassified
Only Key frame	**SIFT**	**75.7%**	**19.9%**	**4.4%**
Movie shots (w/ SPM)	**SIFT**	**81.9%**	**16.8%**	**1.3%**
	OP	**66.5%**	**32.9%**	**0.5%**
	SIFT+OP	72.6%	27.2%	0.3%
	SIFT*w + OP	**84.1%**	**15.9%**	**0.1%**
Movie shots (w/ STPM)	**SIFT**	**84.3%**	**15.6%**	**0.1%**
	OP	**65.5%**	**34.5%**	**0.0%**
	SIFT+OP	72.8%	27.2%	0.1%
	SIFT*w + OP	**86.3%**	**13.7%**	**0.0%**

Figure 9: Experimental results in quality of binary decision

with spatio-temporal pyramid matching (STPM). The precision accuracy of SIFT matching for key frames was 75.7%. However, the precision accuracy was increased to 81.9% by multiple frames of the video shots. The optical flows helped to improve the precision accuracy into 84.1%. The pyramid matching achieved 86.4% of precision accuracy.

6. CONCLUSIONS

In this paper, we addressed the problem of partitioning a video into several video shots and classifying the shots for content-based querying. The shot boundaries are found using a strong classifier learnt from a boosting algorithm on top of weak classifiers. Then, the similarity of video shots is calculated by our spatio-temporal pyramid matching which includes temporal dimension into the matching schema.

As for our experiment, we used sample video clips from sports events. Our experimental results show that the temporal dimension is effective feature to match video shots. We compared several different classification analyses based on used features (such as SIFT and entropy of optical flows), and matching method (such as one key frame, spatial pyramid matching, and spatio-temporal pyramid matching). Our spatio-temporal pyramid matching achieves 86.4% in querying precision of a binary decision.

7. ACKNOWLEDGEMENT

This research was supported by a grant(07KLSGC05) from Cutting-edge Urban Development - Korean Land Spatialization Research Project funded by Ministry of Construction and Transportation of Korean government.

8. REFERENCES

[1] D. a. Adjeroh, M. Lee, and I. King. A distance measure for video sequences. *Vision and Image Understanding*, 85(1/2):25–45, July/August 1999.

[2] R. Ando, K. Shinoda, S. Furui, and T. Mochizuki. A robust scene recognition system for baseball broadcast using data-driven approach. In *6th ACM international conference on Image and video retrieval (CIVR '07)*, Amsterdam, Netherlands, 2007.

[3] P. Chang, M. Han, and Y. Gong. Extract highlights from baseball game video with hidden markov models. In *IEEE International Conference on Image Processing*, 2002.

[4] L. Fei-Fei, R. Fergus, and P. Perona. One-shot learning of object categories. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28(4):594–611, 2006.

[5] D. A. Forsyth and J. Ponce. *Computer Vision: A Modern Approach*. Prentice Hall, August 2002.

[6] Y. Freund and R. E. Schapire. A decision-theoretic generalization of on-line learning and an application to boosting. In *European Conference on Computational Learning Theory*, pages 23–37, 1995.

[7] Y. Gong, W. H. M. Han, and W. Xu. Maximum entropy model-based baseball highlight detection and classification. *International Journal of Computer Vision and Image Understanding*, 96(2):181–199, 2004.

[8] Google. Google video. "http://video.google.com".

[9] K. Grauman and T. Darrell. The pyramid match kernel: Discriminative classification with sets of image features. In *IEEE International Conference on Computer Vision (ICCV)*, Beijing, China, October 2005.

[10] T. C. Hoad and J. Zobel. Fast video matching with signature alignment. In *5th ACM SIGMM International Workshop on Multimedia Information Retrieval (MIR '03)*, Berkeley, California, USA, November 2003.

[11] K. Järvelin and J. Kekäläinen. Ir evaluation methods for retrieving highly relevant documents. In *SIGIR '00: Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval*, pages 41–48, New York, NY, USA, 2000. ACM.

[12] S. Lazebnik, C. Schmid, and J. Ponce. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In *CVPR '06: Proceedings of the 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 2169–2178, Washington, DC, USA, 2006. IEEE Computer Society.

[13] M. S. Lew, N. Sebe, C. Djeraba, and R. Jain. Content-based multimedia information retrieval: State of the art and challenges. *ACM Transactions on Multimedia Computing, Communications and Applications*, 2(1):1–19, February 2006.

[14] B. Li and M. Sezan. Event detection and summarization in sports video. In *IEEE Workshop on Content-Based Access of Image and Video Libraries*, 2001.

[15] D. G. Lowe. Towards a computational model for object recognition in IT cortex. In *Biologically Motivated Computer Vision*, pages 20–31, 2000.

[16] B. Lucas and T. Kanade. An iterative image registration technique with an application to stereo vision. In *IJCAI81*, pages 674–679, 1981.

[17] S. Marchand-Maillet. Content-based video retrieval: An overview, May 2002.

[18] Microsoft. Msn video. "http://video.msn.com".

[19] T. Mochizuki, M. Tadenuma, and N. Yagi. Baseball video indexing using patternization of scenes and hidden markov model. In *IEEE International Conference on Image Processing*, 2005.

[20] M. J. Pickering and S. Ruger. Evaluation of key-frame based retrieval techniques for video. *Computer Vision and Image Understanding*, 92(2/3):217–235, November/December 2003.

[21] H. T. Shen, B. C. Ooi, and X. Zhou. Towards effective indexing for very large video sequence. In *ACM SIGMOD 2005 Conference*, Baltimore, Maryland, USA, June 2005.

[22] H. T. Shen, X. Zhou, Z. Juang, and K. Shao. Statistical summarization of content features for fast near-duplicate video detection. In *15th ACM International Conference on Multimedia (MM '07)*, Augsburg, Bavaria, Germany, September 2007.

[23] X. Tong, Q. Liu, L. Duan, H. Lu, C. Xu, and Q. Tian. A unified framework for semantic shot representation of sprots video. In *7th ACM SIGMM International Workshop on Multimedia Information Retrieval (MIR '05)*, Singapore, November 2005.

[24] A. Ulges, C. Schulze, D. Keysers, and T. Beuel. Content-based video tagging for online video portals. In *3rd MUSCLE ImageCLEF Workshop on Image and Video Retrieval Evaluation*, 2007.

[25] D. Xu and S.-F. Chang. Visual event recognition in news video using kernel methods with multi-level temporal alignment. In *CVPR*. IEEE Computer Society, 2007.

[26] G. Xu, Y.-F. Ma, H.-J. Zhang, and S. Yang. Motion based event recognition using hmm. In *IEEE International Conference on Pattern Recognition*, 2002.

[27] P. Xu, L. Xie, S.-F. Chang, A. Divakaran, A. Vetro, and H. Sun. Algorithm and system for segmentation and structure analysis in soccer video. In *IEEE International Conference on Multimedia and Expo (ICME '01)*, 2001.

[28] Yahoo! Yahoo! video. "http://video.yahoo.com".

[29] YouTube. Youtube - broadcast yourself.

[30] J. Yuan, L.-Y. Duan, Q. Tian, and C. Xu. Fast and robust short video clip search using an index structure. In *6th ACM SIGMM International Workshop on Multimedia Information Retrieval (MIR '04)*, New York, New York, USA, October 2004.

[31] Y. Zhang, X. Zhang, C. Xu, and H. Lu. Personalized retrieval of sports video. In *9th ACM SIGMM International Workshop on Multimedia Information Retrieval (MIR '07)*, Augsburg, Bavaria, Germany, September 2007.