A GENERIC MID-LEVEL REPRESENTATION FOR SEMANTIC VIDEO ANALYSIS

Qing Tang[1], Joo-Hwee Lim[2], Jesse S. Jin[1,3], Haiping Sun[2], Qi Tian[2]

[1]School of Information Technologies, University of Sydney, NSW 2006, Australia
{qtang, jesse}@it.usyd.edu.au
[2]Institute for Infocomm Research, 21 Heng Mui Keng Terrace, Singapore 119613
{joohwee, haiping, tian}@i2r.a-star.edu.sg
[3]School of Design, Communication and I.T., University of Newcastle, NSW 2308, Australia
jesse.jin@newcastle.edu.au

ABSTRACT

This paper presents a generic mid-level representation for efficient semantic video analysis, which adopts a frame-by-frame scheme using P-frames rather than shot-based schemes. Each P-frame is partitioned into an m×n grid (row by column), and each cell is called a 'block'. The representation can bridge the semantic gap and build an intermediate description of video features across frames and blocks. Soccer video is used to showcase the potential of the framework for real video processing. In addition, experiments with tennis video and news video have also been conducted. Results demonstrate the excellent performance of the framework in semantic analysis and also indicate its further potential for automatic video analysis.

1. INTRODUCTION

The large amount of video data raises many research issues, among which the field of video processing focuses on video modeling, video parsing and video retrieval. Despite the significant progress in automated video processing, there are drawbacks in current systems which make them far from users' expectation.

In most video processing approaches, video data are segmented into shots in the low-level feature analysis [1]; and then the frame-to-frame similarity within a shot is exploited to generate compact video representation by key frames [2]. The key frames sequence forms a mid-level representation of the video and high-level analysis is performed over the representation. However, a shot is rather a physical unit than a semantic unit. The information lost in this summarization may impair further analysis.

Currently, most approaches focus on a specific video domain in order to investigate statistical learning structural and semantic meaning. Although some promising results have been reported, it is hard to extend the approach developed for one kind of video to another due to the strong domain knowledge and specialized domain models.

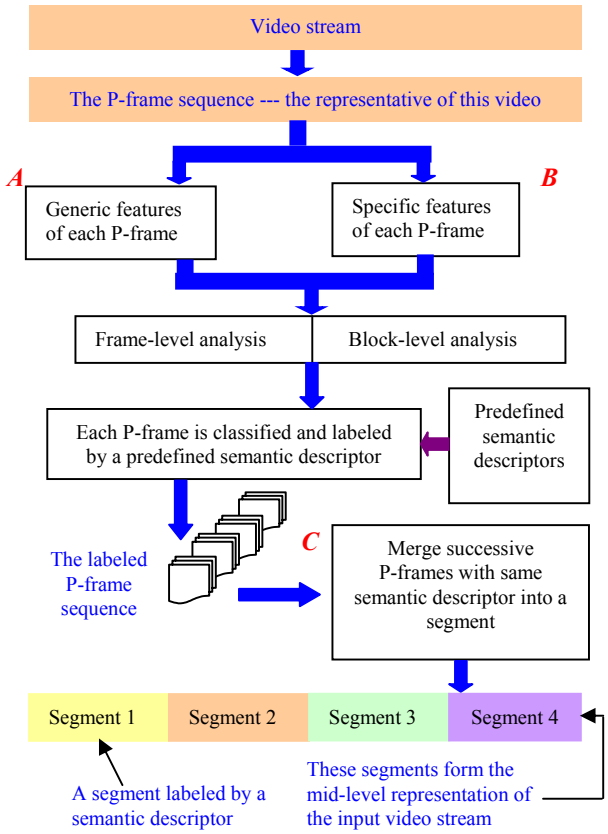

Fig.1. A generic mid-level representation for videos

This paper aims to provide an effective and generic mid-level representation for high-level analysis of various video categories. The main steps of the proposed method

0-7803-8554-3/04/$20.00 ©2004 IEEE.

are shown in Fig.1. In this method, a P-frame sequence is the representative of an input video and the analysis is based on P-frame rather than on shot. Firstly, in order to bridge the gap between low-level features and semantic meanings, we predefined *semantic descriptors* (SD) for each kind of videos according to their structure and the needs for analysis such as event detection in them. Each P frame is partitioned into an m×n grid (row by column), and each cell is called a 'block'. The analyses based on both frame and block basis are performed on all P-frames of the input video stream. Then, each P-frame is classified into predefined categories. Each category is labeled by a SD to indicate its semantic meaning. Hence, the video stream is represented by a set of labeled P-frames. Then, merging process is performed in this set so that successive P-frames with the same SD are gathered into the same segment. Hence, the video stream is converted to a set of semantic labeled segments. This is the mid-level representation of this video stream.

We emphasize three components of proposed method: 1) a novel mid-level representation for semantic video analysis; 2) a frame-by-frame scheme instead of traditional shot-based schemes; and 3) an innovative method of analyzing each P-frame at both frame and block levels.

The rest of this paper is organized as follows: the detailed description of the processing steps in Fig.1 is described in Section 2. In Section 3, to evaluate the effectiveness of this generic method, we depict its application in soccer video as an example; then experimental results and conclusion are given in Sections 4 and 5, respectively.

2. THE PROPOSED GENERIC MID-LEVEL REPRESENTATION

In this section, we first argue the reasons why we do not use the traditional shot-based video processing. Then, our novel mid-level representation is introduced to demonstrate the semantic hints as well as the definitions for soccer, tennis and news videos. How to extract features in the two different level-based analyses are discussed. Finally, how to label each P-frame by a SD is introduced followed by the method used to produce the representation (Step '***C***' in Fig.1) is briefly presented.

2.1. Drawbacks of shot-based analysis

To reduce computational complexity, shots are segmented as a collection of frames sharing the same high-level features, as well as similar low-level features. Most existing segmentation algorithms [1] rely on suitable threshold of differences between successive frames. However, these thresholds are typically highly sensitive to the specific type of video. Furthermore, for high-level semantic video analysis, a shot sometimes contains too many meanings to be effectively adopted as the content unit; or, there is no one-to-one relationship between shot and content unit. For example, the pictures in Fig.2 are selected from a shot of news video. In this shot, the anchorperson finishes reporting two stories. The shot is not a suitable content unit in analysis. Another example is shown in Fig.3. The whole shot includes three areas in the field: penalty boxes of both side and the area between them. Because most important events such as shooting and scoring happened within or around the penalty area, a sequence of frames including penalty box should be considered a semantic segment. It is different from a sequence of frames showing actions around the center circle, which should be regarded as another semantic segment.

With the consideration of designing a generic mid-level representation with applications in different kinds of videos, shot segmentation is not adopted as the traditional first step in this proposed representation. No actual segmentation is done and the analysis is based on all P-frames of a video stream.

Fig.2. A Shot from a news video sequence

Fig.3. Frames selected from a shot of soccer video

2.2. Mid-level representation

We aim to design an innovative mid-level representation to bridge the gap between low-level features and semantic meanings for further video analysis. In [3], Lim defined a new concept called *visual keywords* to delineate images' semantic meanings. Inspired by his method, we reveal that we can define a concept called *semantic descriptor* (SD) to indicate the semantic meaning of video frame or segment. All the SDs defined for certain video type form a *semantic descriptors set* (SDS) for this kind of video.

In order to overcome drawbacks of using shot segmentation, we use a SD to depict a P-frame as shown in Fig.1. After merging procedure (Step '***C***' in Fig.1), a video can be converted to a set of SD labeled segments. Hence, our proposed mid-level representation is obtained in the format of a SD sequence for further semantic analysis.

As presented in Fig.1, we defined some corresponding SDS for different kinds of videos. These sets are stored in a SDS database in the form of type, semantic meanings, and other domain specific description. The definition of

SDS for soccer video, tennis video and news video are listed in Tables 1-3.

2.3. Feature extraction

There are many well-developed technologies in video processing and computer vision communities. However, we have to consider the computational efficiency when choosing those technologies; and also, a video cannot be treated as a sequence of still images only without considering information in temporal dimensions. We have to find a balance between speed and analysis results.

Table 1. Defined SDS for soccer video

SDS	Semantic meanings	Description
AD	Audience	Far view of audience
FM	Fast movement toward a penalty box or fight for ball control; A break happened between two penalty boxes.	Far view of whole field (goal post not visible)
FP	Move inside or outside a penalty box; Players are waiting for free kick, corner kick or break.	Far view of half field (goal post visible)
MB	Actions such as chasing the ball between players; Players are waiting for free kick or corner kick.	Mid-range view (whole body visible)
CP	Close up	Close-up of a player, referee, coach, goalkeeper without field appearance
GP	Free kick, corner kick, goal, shoot or goal kick	Goal post in close-up view
Player (s)	Player fouled, missed a chance or took over a free kick	Mid-range or close-up of a player

Table 2. Defined SDS for tennis video

SDS	Semantic meanings	Description
AD	Audience	Far view of audience
CV	Play	Court view
MB	Break	Mid-range view
FW	Begin to play	Far view of whole court
CP	Break or serve	Close-up view

Table 3. Defined SDS for news video

SDS	Semantic meanings	Description
ACs	Read news	Anchor person (s) or correspondent
BK	Break of the news	Commercial, weather report
ST	News	News story (anchor person not visible)

To find the balance, we propose analysis at two different levels, frame-level and block-level, shown as '***A***' and '***B***' in Fig.1. The frame-level analysis means the features are extracted to delineate the character of a frame. For example, color histogram of one P-frame is for frame-level analysis. Block-level analysis means some features are extracted from one or several blocks and statistically learned to describe the character of blocks. For example, edge density of a block is a feature for block-level analysis.

We have to achieve another balance between the commonness of features and utility of domain knowledge to improve analysis results. We choose a few features commonly used in all kinds of videos while a few specific features according to certain video domain knowledge. The extracted generic features are listed in Table 4.

Table 4. Extracted generic features

	Frame-level	Block-level
Color	Histogram of a frame; Dominant color.	Histogram of a block; Dominant color; Number of colors in a block; Number of colors of blocks in a row.
Motion	Direction histogram of a frame; Magnitude of a frame.	Direction histogram of a block; Magnitude of a block.
Texture	Histogram of gradient magnitudes of a frame; Histogram of gradient orientations of a block.	Histogram of gradient magnitudes of a block; Histogram of gradient orientations of a block.

2.4. Labeling P-frames by SDS

Szummer et al. in [4] presented a method to classify images into two categories: indoor and outdoor. In this method, each frame is divided into 4 by 4 blocks. Each block is labeled with either 'in' or 'out' with respect to its features analysis. Then the image is labeled 'indoor' or 'outdoor' according to all 16 labels within an image.

Inspired by their approach, we define a few basic and meaningful *block descriptors* (BD) to depict each block of a frame so that we can deduce the semantic meaning of this frame. Using extracted features described in Section 2.3, Support Vector Machine (SVM) classifies and labels each block of a frame by a BD.

This is just auxiliary for labeling of frames, and the relationship between the BDs and the SDS must be clear before using this set.

3. APPLICATION IN DIFFERENT VIDEO DOMAINS

In this section, we illustrate the effectiveness of our generic framework with application in semantic video analysis. We choose the soccer video as an emphasis. The analyses of tennis video and news video are the same with slight modification according their domain knowledge.

The SDS for soccer video is listed in Table 1. There are no specific features extracted for the block-level analysis, but only generic features presented in Table 4. At the frame-level, we detected goal post in close-up view

by searching bold goal bar or goal net and detected goal poles in far view by using Hough Transform to detect two white parallel lines. The algorithm is similar to that discussed in [5]. We defined four BDs: A for 'Audience', G for 'Ground', B for 'Body' and U for 'Unknown', to represent each block. SVM is used to distribute a BD to each block. We also defined a mapping relationship between SD and SL. An example of this relationship is shown in Fig.4. Therefore, we can infer the SD of each P-frame from its 12 block BDs. Information obtained from frame-level feature analysis helps us further modify the inference results.

After the steps described above, the mid-level representation for the video is produced.

A	A	A	A	A	A
A	G	G	U	U	A
G	G	G	G	G	A
G	G	G	G	G	G

This P-frame can be labeled as FP

The BD labeled blocks in a P-frame can infer this is a FP frame

Fig.4. A SD and SLs mapping example

4. EXPERIMENTAL RESULTS

For the differences of categories and domain knowledge, we produced three different systems to produce mid-level representations for soccer, tennis and news videos. The results are shown in Tables 5-7 respectively.

Table 5. SDs Detection results in soccer videos

	Ground Truth	Output	Correct	Accuracy
AD	62	68	56	82.4%
FM	659	671	584	87.0%
FP	507	497	422	83.2%
MB	314	286	218	76.2%
CP	345	394	334	84.7%
GP	53	59	46	77.9%
Player	296	273	202	74.0%

Table 6. SDs detection results in tennis video clips

	Ground Truth	Output	Correct	Accuracy
AD	14	16	13	81.3%
CV	37	37	32	86.5%
MB	19	18	13	72.2%
FW	6	6	5	83.3%
CP	26	25	21	84.0%

Table 7. SDs detection results in news video clips

	Ground Truth	Output	Correct	Accuracy
ACs	38	35	30	85.7%
BK	6	5	4	80.0%
ST	38	42	34	81.0%

The test data set includes: 1) 5 games from FIFA2002, in total of 450 minutes without commercial for soccer video test; 2) 90 minutes clips from Australia Open 2001 for tennis video test; and 3) 60 minutes news video clips from different broadcast stations.

When testing, the output is correct if the offset of an output segment is less than 10% of the length of the ground truth at both the start end and the rear end.

In order to test the effectiveness of the proposed mid-level representation, we defined some 'events' in these three kinds of video and use cinematic features similar to those in [5] to detect them. The results are shown in Table 8. From the results we conclude that the proposed mid-level representation is generic and effective.

Table 8. Results for event detection in these videos

	Corrected detected	Miss	False alarm
Goal in soccer video	9	2	2
Fighting for ball possession (similar to rally in tennis) in soccer video clips	31	6	7
First serve scored in tennis video clips	11	4	5
Weather reports in news video clips	3	1	1

5. CONCLUSION

This paper presents a generic mid-level representation for semantic video analysis. A video is converted into a representation in form of a SD sequence. The generic mid-level representation, frame-by-frame analysis instead of shot-based analysis and two-level analysis of each P-frame are the main contributions of this paper. Soccer, tennis and news videos are used to test the approach. Experiments have shown robust performance in semantic analysis.

Audio and text features will be added to improve its performance. Also, more varieties of videos will be used to test the robustness.

6. REFERENCES

[1] I. Koprinska, S. Carrato, "Temporal video segmentation: a survey", Signal Processing: Image Communication, 2001.

[2] M. M. Yeung, B-L. Yeo, W. Wolf, and B. Liu, "Video browsing and scene transitions on compressed sequences", in Symp. Electronic Imaging: Science and Technology: Storage and Retrieval for Image and Video Databases IV. 1997, IS&T/SPI97.

[3] J.H. Lim. "Building Visual Vocabulary for Image Indexation and Query Formulation", Pattern Analysis and Applications (Special Issue on Image Indexation), 2001.

[4] M. Szummer, and R. W. Picard, "Indoor-outdoor Image Classification", IEEE International Workshop on Content-based Access of Image and Video Databases, 1998.

[5] A. Ekin, A. M. Tekalp and R. Mehrotra, "Automatic soccer video analysis and summarization", IEEE Trans. On Image Processing, June 2003, in press.