DETECTION OF THE HIGHLIGHTS IN BASEBALL VIDEO PROGRAM

Huang-Chia Shih and Chung-Lin Huang
Institute of Electrical Engineering
National Tsing Hua University
HsinChu, Taiwan,
clhuang@ee.nthu.edu.tw

ABSTRACT

This paper introduces a highlight detection method that consists of (1) inference scheme for semantic feature extraction, and (2) a semantic highlight detection scheme using a Multi-level Semantic Network (MSN). MSN is basically a Bayesian Belief Network (BBN) which consists of the low-level feature nodes connected to the mid-level semantic nodes and finally the high-level semantic nodes. A probabilistic structure can be combined by visual features and can be applied on highlight detection. Satisfactory results have been shown in the end of this paper.

1. INTRODUCTION

Recently, video indexing for sports video has become an active research topic. Sport programs are not persistently live exciting and the program highlights seem more interesting to the TV audience. The highlights indicate certain kinds of significant offending or defending events that may cause the viewer's interest or preference in the sports video program. A baseball program usually comprises almost 80% non-highlight scenes. This paper proposes a highlight detection algorithm using Bayesian Belief Network (BBN) for baseball video program.

The pioneer work[1] using the Bayesian architecture for video content characterization and analysis provides a potential tool for accessing and browsing video database on a semantic basis. The BBN is a directed acyclic graph, which has been proved to be an effective knowledge representation and inference engine in artificial intelligence and expert system. Ferman *et al.* [2] employed Hidden Markov Model (HMM) and BBNs at various stages to characterize the content domain and extract the relevant semantic information. BBN has been applied for video scene classification[3], object detection and tracking [4], and semantic interpretation [5].

Chang *et al.* [6] describe a similar approach for highlights extraction directly from four statistical models based on HMM. Any one type of four highlights is corresponding to a HMM individually. But the accuracy of highlight detection is very sensitive to the precision of shot classification phase. Amhet *et al.* [7] introduce shot identification framework and goal detection for soccer game. Their method seems not be objective and the goal detection is not only too trivial but restricted to the accuracy of shot type classification. Li *et al.* [8] propose a good concept to model sports video using events. First, they detect some crucial scene cut points such as setup action, live action, threshold of close-up view, and replay. Based on the temporally relationship of scene shots, their method can fetch out the location of key events.

This paper presents a multi-level Semantic Network (MSN) for highlight detection in the baseball video. Based on the low-level information and the inferring processes, the MSN will infer the high-level semantic of the video. In sports, the highlights that unfolded are governed by the rules. Hence they contain a recurring temporal structure. The rules of production of such videos have also been standardized. For example, in baseball videos, there are only a few recurrent views, such as pitching, close-up, home plate, battering, crowd etc. Our system is designated for understanding the baseball video, which can be easily modified for other sports videos. Different from the previous researches, this paper proposes a MSN that uses the low-level image information for semantic inference to find the highlights in the sports video program.

2. THE PROPOSED FRAMEWORKS

Here, our approach that supports an inference of unobservable concepts based on their relevance with the observable evidences. Given low-level evidences, the statistical model-based classifiers and Multi-level Semantic Network (MSN) may infer certain high-level concepts. We develop several SNs to model the different semantic events in the baseball video.

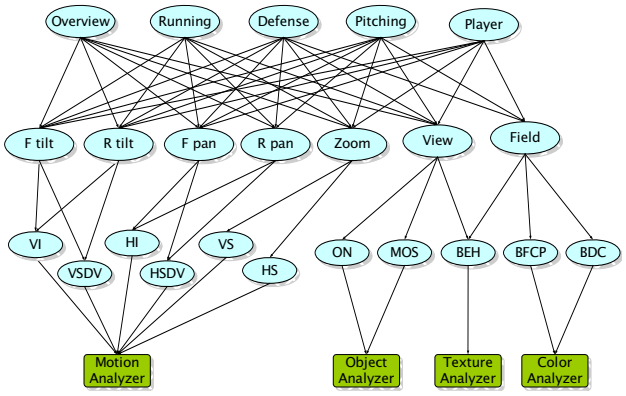
Having applied the BBN training procedure on the MSN, we may use the MSN to interpret the semantic meanings of different events in the video. Given a video in a specific domain, our system may extract the low-level evidences and then translate the input video into high-level semantic meaning. Specific domains contain rich spatial and temporal transitional structures. MSN is a modified the BBN structure which can be applied for the highlight detection in the baseball video.

The semantic concept *View* is modeled by a MSN connecting several low-level evidences such as *Object-Number* (ON), *Main-Object-Size* (MOS) and *Edge Texture*. Inferences from two image analyzers: texture and object analyzer provide these low-level evidences. The ON and MOS are significant evidences to infer the *View*. If the evidence of a large ON is obvious then it may strongly support the possibility that node *View* favors the distant-view rather than the close-up view. Similarly, if the evidence of a large MOS is obvious then the node *View* will indicate a close-up view rather than a distant-view.

On the top of the MSN (see Figure 1) are the root nodes

representing the certainty of the five different scene categories: *Overview, Running, Defense, Pitching* and *Player*. The upper level of the MSN may infer the highest-level concept of the root node from the input video sequence. Each input video may activate more than one root node (with high certainty after BBN inference). Each root node is connected to several mid-level nodes representing the semantic concepts.

Here, we use seven mid-level nodes to represent semantic concepts such as *fast tilting, regular tilting, fast panning, regular panning, zooming, view* and *field*. These semantic concepts originate from the instinct response of the viewers of the sports program. Each root node represents the category of a certain video shot. The linkage characteristics of the MSN are also manually determined based-on their relationship, and the probabilities of these links can be obtained by the BBN training procedure[9,10].



NOMENCLATURE

F tilt : Fast Tilting
R tilt : Regular Tilting
F pan : Fast Panning
R pan : Regular Panning
Zoom : Zooming
VI: veritcal Intercept
VSDV: Vertical Sum of Displacement Vector
HI:Horizontal Intercept
HSDV:Horizontal Sum of Displacement Vector
VS:Vertical Slope
HS: Horizontal Slope
ON: Object Number
MOS: Main Object Size
BEH: Background Edge Histogram
BFCP: Background Field Color Percentage
BDC: Background Dominate Color

Figure 1. The MSN of the event interpretation.

The high-level semantic, which is considered as indirect aggregations of lower level information, may also be represented by the MSN provided that they can be inferred directly to the semantic of input video. Figure 1 illustrates the MSN for video event interpretation of the baseball video. The overall structure consists of three layers: the category layer, the mid-level semantic layer, and the low-level feature layer. In Figure 1, VI, HI, VS and HS are the parameters of the linear regression in motion analyzer.

Highlights are found in the video clips indicating the occurrence of certain particular occasion or specific circumstance. The main concern is the choice of the variables that are necessary to demonstrate a highlighted clip from the human visual conception. The MSN for highlight detection is illustrated in Figure 2. The highlight will be found in the scenes of *defense, hits, grounder line* and *cloud buster*, but cannot be found in the scenes of *pitching, player close-up* and *overview*.

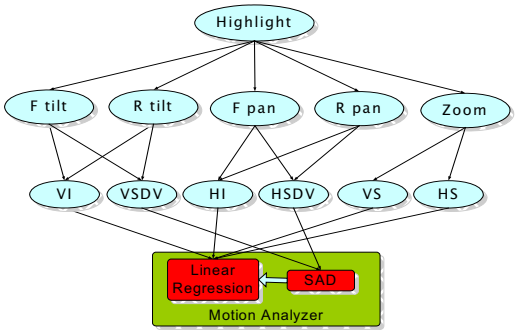


Figure 2. The MSN for highlight detection.

3. LOW LEVEL INFORMATION EXTRACTION

The video browsing, summarization, and retrieval are based on the low-level information analysis and the high-level inference of the digital video. To extract the low-level evidences, we need to analyze the object motion, the colors, the textures and the diversified camera motion.

1) **Texture Analysis.** The edge histogram descriptor (EHD) is also applied to describe the spatial distribution of edges, which is useful for matching image even when the underlying texture is not homogeneous. The edge information is categorized into five classes: vertical, horizontal, 45° diagonal, 135° diagonal, and isotropic. To select the dominant component of edge direction treated as the texture feature.

2) **Color analysis.** Color has been important information embedded in the video. Here, we use the CIE-YUV color space to analyze the color information of which can be used as a low-level feature. A dominant color and its percentage value will be calculated.



Figure 3. The moving objects segmentation (the red block depicted the major object).

3) **Moving Object Segmentation**. The moving object region and background region are separated before the feature extraction such as texture, color and motion information. Here, we assume that the moving objects in complex background are somehow identifiable by their edge boundaries. Usually, the edge information is too noisy to be applicable for computer vision system, and most of the edge information is redundant. We assume that the objects

are moving, and the background scene is complex but stationary. The results of the segmentation processes are illustrated in Figure 3. The segmentation process is vulnerable due to the unpredictable edited scene. For instance, the last picture in Figure 3 shows an object is segment into to object because the caption separates the object into two.

4) **Motion analysis.** The motion information consists of three cases: *(1)zooming, (2)panning, (3)tilting*. To test the effect of camera motion for video sequence, we replace the time consuming calculation of 2-dimensional $m \times n$ picture elements with that of two one-dimensional vectors. This is made possible by mathematically operating the luminance values of vertical and horizontal lines as the characteristic values of x and y direction respectively. There are two steps for motion analysis: (1) find the displacement vector that makes the minimum SAD (Sum of Absolute Difference) values; (2) use a simple linear regression to obtain the Displacement Characteristic (DC) curve as shown in Figure 4.

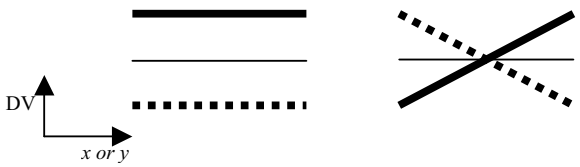


Figure 4. The DC curve. (a) Panning /tilting to the left/down (solid line) and to the right/up (dotted line). (b) Zoom in (solid line) and zoom out (dotted line)

The kth observation z_k of parameter node λ for motion analyzer obtained from the motion analyzer is described as

$$z_{\lambda}^{k} = \sum_{n=k}^{k+cu-f_o} \lambda^{(n)} \Big/ (cu - f_0 + 1) \qquad (1)$$

where λ can be HS, VS, HI, VI, HSDV, or VSDV, f_0 is the number of frames required for one observation, *cu* is the coding unit (frame), *HSDV* and *VSDV* are defined as

$$HSDV^{(n)} = \sum_{p=1}^{(W-N)/n_0} DV_p^{horizontal,n:n+1}$$

$$VSDV^{(n)} = \sum_{p=1}^{(H-N)/n_0} DV_p^{vertical,n:n+1} \qquad (2)$$

where $H \times W$ is frame size, N is shifting index; n_0 indicates the slice windows size. The parameters generated by the object, texture and color analyzer are defined as follows:

$$z_{\upsilon}^{k} = medium(\upsilon_k : \upsilon_{k+cu-1}). \qquad (3)$$

where υ can be ON, MOS, BEH, BDC, or BFCP.

4. BAYESIAN BELIEF NETWORK

For semantic understanding of the input video sequence, we demonstrate that BBN can be applied to the specific video to extract the corresponding semantic contents. BBN has been proved to be an effective statistical model for knowledge representation and inference. BBN is a direct acyclic graph representing the causal/relevance dependencies between variables, which are represented with the conditional probabilities. In BBNs, variables are used to represent events and/or objects in the world. We may integrate prior information about dependencies between variable and propagate the impact of evidence on the probabilities of uncertain outcomes. Here, we apply BBN to modeling our MSN network. In BBN, direct arcs between variables represent conditional dependencies.

There are two types of computations performed with BBNs: *belief updating* and *belief revision*. Belief updating concerns the computation of probabilities over variables, while belief revision concerns finding the maximally probable global assignment. More formally, we assume W is the set of all variables in BBN, and e is the given evidence which represents a set of instantiations made on a subset of W. Any complete instantiations to all the variables in W which is consistent with e will be called an explanation or interpretation of e. Our goal is to find an explanation w^* such that $P(w^* \mid e) = max\ P(w \mid e)$ where w^* is called the *"*most-probable explanation*"*.

Belief updating on the other hand is interested only in the marginal probabilities of a subset of variable given the evidence. It is applied to determine the best instantiation of a single variable given the evidence. According to Bayes' rule, the posterior probability can be expressed by the joint probability, which can be further expressed by conditional probability and prior probability as

$$P(S \mid E) = \frac{P(S, E)}{P(E)} = \frac{P(E \mid S)P(S)}{P(E)} \qquad (4)$$

where S denotes semantic concept and E denotes evidence. Assume a BBN for a set of variables $X=\{x_1, x_2, ..., x_n\}$, a set of P denotes local probability distributions associated with each variable. The network structure S is a directed acyclic graph. The nodes in S are in one-to-one correspondence with the variables X, x_i denotes both the variable and its corresponding node, and Pa_i denotes the parents of node x_i in S as well as the variables corresponding to those parents. Using the *chain rule*, we may express the joint probability distribution for **X** as

$$P(X) = P(x_1, x_2, ..., x_n) = \prod_{i=1}^{n} P(x_i \mid Pa_i) \qquad (5)$$

So, a complicated joint probability distribution can be reduced to a set of conditional probability and a prior probability.

In the training phase, to evaluate the class conditional density function (*i.e.,* conditional probabilities, posterior probabilities), we apply the EM (Expectation- Maximization) algorithm for batch learning [9, 10]. Batch learning is an "off-line" learning, and it is used when we want to generate our conditional probability tables in the knowledge base from the training data stored in a database. Based on different cases given in the training data, the EM-algorithm computes the conditional distribution probability for each node.

Having constructed and trained a MSN, we need to determine various probabilities of interest from the model by

inference procedure, which gives the observations and evidences (*i.e.,* low-level media feature).

5. EXPERIMENTAL RESULTS

In the experiment, we emphasize that the measurement of recognition rate is based on a coding unit rather than the unit of video shot. The baseball video is selected from six different baseball TV programs. Each video shot consists of a sequence of image frames and it indicates different semantic content, moreover, each shot may consist of different number of image sequences, some of which are used for training and the others are used for testing. The size of each full color image frame is 352×240, and its frame rate is 30 frames per second.

The image pre-processing, low-level feature extraction procedure and moving object segmentation are based on the decompressed MPEG video. The blocking artifacts of the input video reduce the efficiency of the low-level information extraction. In the beginning, the states of the nodes of the MSN were specified manually. Every node consists of several discrete states. Each of the identified lowest level features must be assigned to a certain numeric value. The others including mid-level semantic nodes and the nodes of category layer are assigned two states equally. It may be assigned either *"True"* or *"False"*. For the semantic *View*, it may be assigned "*close-up"* and *"distant"*.

Table 1. The results of highlight detection

	Highlights	Non-highlights
# of Training Unit / # of Testing Unit	523/389	477/430
# of Correct/Reject/Mismatch	337/0/52	379/2/49
Precision	86.6%	88.6%
Recall	86.6%	88.1%

Using MSN, we may easily detect the highlight as shown in Table 1 as well as the mid-level semantics as shown in Table 2 that performed by the precision and recall. The precision and recall are computed as:

$$precision = (R \cap C) / R \quad (6)$$

$$recall = (R \cap C) / C \quad (7)$$

where R denotes set of coding units recognized as highlight or corresponding mid-level semantic units, and C represents the relevant set of correct units.

Here we assume coding unit cu=6 frames. In the experiments, We have selected about 1100 video shots from six baseball TV programs of Major League Baseball (MLB), three baseball TV programs of Chinese Professional Baseball League (CPBL), three baseball TV programs of Japan Professional Baseball League (JPBL) and four baseball competition of world cup in 2002. Totally, we have 3000 coding units for training and 1176 coding units for testing.

Table 2. The results of the mid-level semantics.

	Field	View	Zooming	Fast panning	Regular panning	Fast tilting	Regular tilting
Precision	79.6%	76.7%	86.6%	89.4%	75.8%	99.4%	87.0%
Recall	78.8%	76.7%	86.2%	89.4%	75.8%	99.0%	86.5%

6. CONCLUSIONS

This proposed technique can be extended to more general interesting sports programs. Experiments show that the results of semantic extraction and highlight detection are acceptable. Our low-level-feature-based highlight detection algorithm has three limitations: (1) memory requirements due to a large amount of conditional probability need simultaneously applied for computation, (2) long latency time that makes it unsuitable for real-time video analysis, (3) high-level knowledge such as common sense and human perceptual information can not be encoded in terms of the low-level features.

7. REFERENCES

[1] N. Vasconcelos and A. Lippman, "Bayesian modeling of video editing and structure: semantic features for video summarization and browsing," IEEE ICIP, Chicago, 1998.

[2] A. M. Ferman and A, M, Tekalp, "Probabilistic analysis and extraction of video content," IEEE ICIP Tokyo, Japan, OCT., 1999.

[3] S. F. Chang and H. Sundaram, "Structural and Semantic Analysis of Video," IEEE ICIP, Vancouver, Sept. 2000.

[4] J. Luo, A. E. Savakis, S. P. Etz, and A. Singhal, "On the application of Bayes Networks to Semantic Understanding of Consumer Photographs," IEEE ICIP, Sept. 2000.

[5] A. Ekin, A. M. Tekalp, and R. Mehrotra, "Extraction of Semantic Description of Event using Bayseian Network," IEEE, ICIP 2001.

[6] P. Chang *et al.* "Extract Highlights from Baseball Game Video with Hidden Markov Models". In IEEE ICIP, 2002.

[7] E. Ahmet and A. M. Tekalp, "Shot Type Classification by Dominant Color for Sports Video Segmentation and Summarization", IEEE ICASSP 2003 Hong-Kong, April, 2003.

[8] B. Li *et al.* "A General Framework for Sports Video Summarization with its Application to Soccer", IEEE ICASSP 2003, Hong-Kong, April, 2003.

[9] Finn V. Jensen, "An Introduction to Bayesian Networks", Springer-Verlag, 1996.

[10] T. K. Moon, "The Expectation-Maximization Algorithm", IEEE Signal Processing Magazine, Nov 1996, pp 47-60.