A SEMANTIC NETWORK MODELING FOR UNDERSTANDING BASEBALL VIDEO

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

ABSTRACT

The exploitation of semantic information in videos is difficult because of the large difference in representations, levels of knowledge and abstract episodes. Traditional image/video understanding and indexing is formulated in terms of low-level features describing image/video structure and intensity, while high-level knowledge such as common sense and human perceptual knowledge are encoded. This paper attempts to bridge this gap through the integration of image/video analysis algorithms with multi-level semantic network to interpret the baseball video.

1 INTRODUCTION

Recently, a large amount of digital media including image, audio, video, streaming video clips, panorama images and 3D graphics have been created. We need a flexible and scalable way to manage the mass media of which the digital video has been widely accepted as the most accessible media. The MPEG-7 has tried to standardize the media access methods based on its content. The video indexing and retrieval is a useful query tool for us to access the media, which consists of automatic classification, summarization and understanding of the video.

Traditionally, the video indexing and retrieval researches have focused on the paradigm of query-by-example (QBE) [1,2]. The ability to query with key-words or key-phases (semantic) instead of examples has motivated the semantic video indexing. However, the difficulty in such a system that supports semantic retrieval using keyword lies in the gap between low-level media features and high-level concepts. Recently, there have been some efforts to bridge the gap, Naphade et al. [3] propose a novel probabilistic framework (multijects and multinets) for semantic indexing and retrieval in digital video.

Vasconcelos et al. [4] introduce the Bayesian architecture for content characterization and analyze its potential as a tool for accessing and browsing video database on a semantic basis. The Bayesian Belief Network (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 [5]. Ferman et al. [6] employ Hidden Markov Model (HMM) and Bayesian Belief Networks (BBNs) at various stages to characterize the content domain and extract the relevant semantic information. Chang et al. [7] develop a classification scheme based on BBNs, which models the interaction of multiple classes at different levels of multi-media. A Bayesian network based method for semantic object extraction in images was also proposed for object detection and tracking [8] and semantic interpretation [9].

Similar to [4,7~9], we propose a multi-level Bayesian Belief Network (BBN) for event interpretation in the baseball video. The input video contains some rich low-level information such as spatial and temporal information. Based on the low-level information and the inferring processes, the BBN will infer the high-level semantic of the video. Different from the previous researches, this paper proposes a semantic-based multi-level Bayesian Network that can be used to bridge this gap between the low-level image information and high-level semantic meaning through BBN inference engine.

2 EXTRACT THE LOW-LEVEL FEATURE

For any video retrieval, summarization or categorization problem are always based on the low-level information analysis and the high-level inference of the digital video database. For video understanding, it is important to extract the low-level evidences, which consist of the object motion, colors, textures and the panning motion of the camera.

1) **Scene Change Detection.** To represent a sequence of frames captured from a unique and continuous record from a camera. Adjacent frames of the same shot exhibit temporal continuity.

2) **Texture.** The Gray-Level Co-occurrence Matrix (GLCM) is used to capture the spatial dependence of gray-level region. 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 [11].

3) **Color.** We use the CIE-YUV color space to analyze the color information, which can be used as a low-level feature. A dominant color and its percentage value will also be calculated.

4) **Motion.** The motion information consists of two cases: (i) zooming (ii) panning. To test the effect of zooming and panning for video sequence, we apply the method mentioned in [10]. They replaces the time consuming calculation of 2-dimensional $m \times n$ picture elements with that of two one-dimensional vectors.

5) **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 image analysis 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 1.

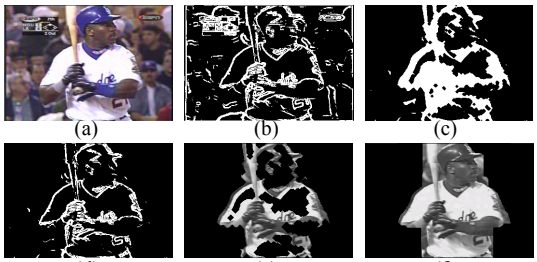


Figure 1. The segmentation (a) original video; (b) edge image (c) after accumulated + closing operation image; (d) result of b AND c; (e) after closing d and rebuild; (f) after noise removal + region growing.

3 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. Bayesian Belief Network (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.

BBN process has been proved to be a powerful mechanism to model the incomplete data and the reasoning in terms of some quantity measurements. In BBN, direct arcs between variables represent conditional dependencies. When all the parents of a given variable A are instantiated, that variable is said to be conditionally independent of the remaining variables, which are not descendants of A.

Assume a Bayesian network for a set of variables $\mathbf{X}=\{x_1, x_2, \ldots, x_n\}$, a set of local probability distributions are 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 $\mathbf{Pa}_i$ 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

$$P(\mathbf{X}) = P(x_1, x_2, \ldots x_n) = \prod_{i=1}^{n} P(x_i \mid \mathrm{Pa}_i) \qquad (1)$$

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

Inference, or model evaluation, is the process of updating probabilities of outcomes based upon the relationships in the model and the evidence known about the situation at hand. When a Bayesian model is constructed, the user applies evidence about recent events or observations. This knowledge is applied to the model by "instantiating" or "clamping" a variable to a state that is consistent with the observation. Instantiating occurs when the state of a variable is known (i.e., hard evidence). Then the mathematical mechanics are performed to update the probabilities of all the other variables that are connected to the variable representing the new evidence.

4. THE PROPOSED FRAMEWORKS

Here, we propose a modeling framework, which supports an inference of unobservable concepts based on their relevance with the observable evidences. Given evidences as the input, the statistical model-based classifiers and Semantic Network (SN) may infer certain high-level concepts. We develop several SNs to model the different semantic events in the baseball video (see Figure 2).

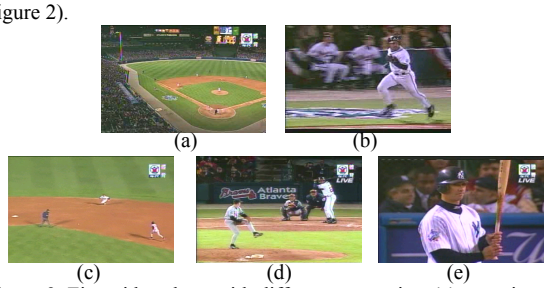


Figure 2. Five video shots with different semantics; (a) overview, (b) runner snapshot, (c) defending view, (d) pitching view; (e) batter snapshot.

We apply the BBN training procedure on the SNs, and then use the SNs 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. For example, in baseball videos, there are only a few recurrent views, such as pitching, close-up, home plate, battering, crowd etc. Here, we develop several SNs that are used to model the mid-level semantic of baseball video such as *view*, *field*, *zooming*, *regular-panning*, *fast-panning*.

The basic level of the framework consists of several different image analyzers. For instance, the object analyzer finds the existence of the main object, the objects' sizes and number, whereas the texture analyzer describes the background in terms of the texture entropy, and edge histogram. Based on the extracted low-level information, the SN "*View*" describes the input video as *distant view* or *close-up view*, whereas the SN *"Field"* infer the semantic concept of the input video as *infield* or *outfield*.

In Figure 3, the semantic concept *View* is modeled by a SN connecting four low-level evidences: *Object-number, Main-object-size, Texture-entropy, Edge-histogram*. The latter two evidences are denoted by two nodes, of which the corresponding states indicate the certainty of the specific texture information obtained from *texture-analyzer*. These low-level evidences are connected to some image analyzers, such as *texture-analyzer, object-analyzer* and *color-analyzer*. The image analyzers are described by square blocks, which provide the low-level information.

The *Object-Number* and *Main-Object-Size* are significant evidence to infer the *View*. If the evidence of a large *Object-Number* 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 *Main-object-size* is obvious then the node *View* will indicate a close-up view rather than a distant-view.

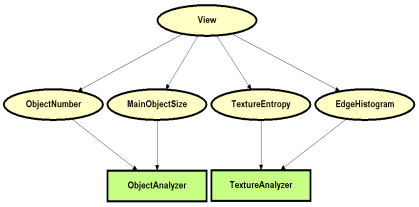


Figure 3. A SN for the semantic concept *"View"*.

Figure 4 illustrates another SN that may be used to infer the mid-level semantic *Field*. The *texture-analyzer* provides low-level information to two low-level semantic nodes, *edge-histogram* and *texture entropy*, which are shared by SN *Field* and SN *View*. The *color analyzer* may provide two new low-level evidences: *Background-Dominant-Color* (BDC) and its percentage value (BDC_P) which are required for inferring the mid-level semantic *Field*. The certainty of either outfield or infield is described by the status of the node *Field*. The image analyzer checks every frame of the video sequences. If the BDC indicates toward green color, then *Field* favors the outfield rather than the infield.

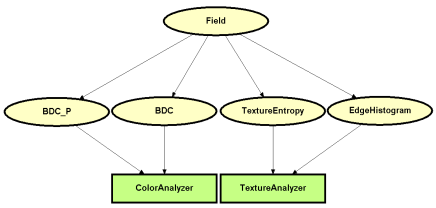


Figure 4. A SN for the semantic concept *"Field"*.

The mid-level semantic concept related to motion activity is described by the SN shown in Figure 5. The displacement vector (DV) information can be used for SN to demonstrate the certainty of three different mid-level semantics: fast panning, regular panning and zooming.

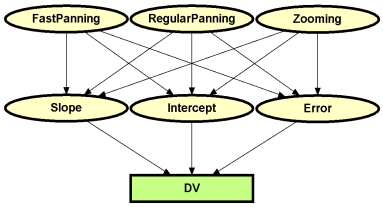


Figure 5. A SN for the semantics concepts *"fast-panning", "regular-panning" and "zooming"*.

On the top of the multi-level BBN hierarchy are the root nodes representing the certainty of the six different categories: *Event-Occur, Overview, Runner, Defense, Pitching* and *Batter*. The upper level of the multi-level network is shown in Figure 6 from which we 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.

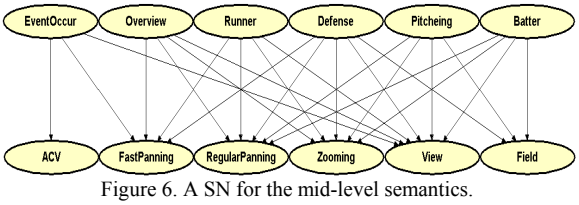


Figure 6. A SN for the mid-level semantics.

Here, we use six mid-level nodes to represent semantic concepts such as *ACV (Aloud Cheering Voice), 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 SN are also manually determined, and the probabilities of these links can be obtained by the BBN training procedure.

The high-level semantic, which is considered as indirect aggregations of lower level information, may also be represented by the SN provided that they can be inferred directly to the semantic of input video. Figure 7 illustrates the SN 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.

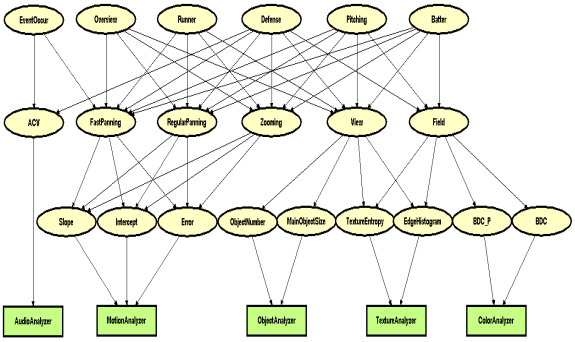


Figure 7. An overall SN for baseball video.

5. IMPLEMENTATION

To apply the BBN technology to our multi-level SN for baseball video semantic event interpretation, we need to consider the following steps:

1. Formulate problem in terms of creating a set of variables representing the distinct elements of the situation being modeled.
2. For each such variable, assign the set of mutually exclusive states or outcomes that it may generate.
3. Settle the causal dependency relationships between the variables. This involves creating arcs (lines with arrowheads) linking from the parent (influencing) node to the child (influenced) node.
4. Assess the numeric probabilities for each variable and arc. Based on the case set given in the training data as well as using the gradient-descent algorithm to compute the conditional distribution probability for each node.

Since the conditional probability and the prior probability for each node are known, we can utilize the BBN model to understand and classify our video sources. The input video is analyzed by several image analyzers: such as *motion analyzer, object analyzer, texture analyzer, and color analyzer.* These analyzers extract the lowest-level features as the input evidence to BBN. Using the evidence propagation procedure from the low-level feature layer to the mid-level semantic layer such as *view* and *field*, we may further infer the high-level semantic category of the video sequence.

6. EXPERIMENTAL RESULTS

In the experiment, we emphasize that the measurement of recognition rate is based on the frame unit rather than the unit of video shot. We have 1100 video shots 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 BBN-based video understanding system may extract the mid-level semantic meanings given some low-level evidences, and it can be used to find the category of testing video shot. Table 1 demonstrates the performance for extracting the mid-level semantic feature including *Field, View, Zooming, Fast panning* and *Regular panning*.

Table 1. Mid-level semantic feature

Mid-level Semantic feature	Training data (frames)	Testing data (frames)	Success rate (%)
Field	12,819	4,161	85.77%
View	17,995	7,777	80.01%
Zooming	11,106	30,238	82.20%
Fast Panning	10,429	27,592	79.24%
Regular Panning	9,724	31,183	74.82%

We use some testing sequences for each category and do the experiments for all the testing sequences, which belongs to the same category. We find that it is categorized correctly if the certainty score of the corresponding node is larger than 50%. However, we may also find that the scores of the other nodes may also be larger than 50%, which are incorrect, and it is called the false alarm. The testing results of detection accuracy and false alarm rate of each category are shown in Tables 2-1 and 2-2.

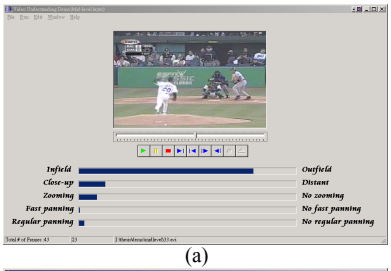
Table 2-1. Testing result for category level.

Input class / Recognized class	A	B	C	D	E
Overview (A)	56	0	1	1	0
Runner (B)	32	52	1	0	4
Defense (C)	13	4	113	3	8
Pitching (D)	1	6	1	66	7
Batter (E)	4	45	0	7	58
Total sequences	106	107	116	77	77

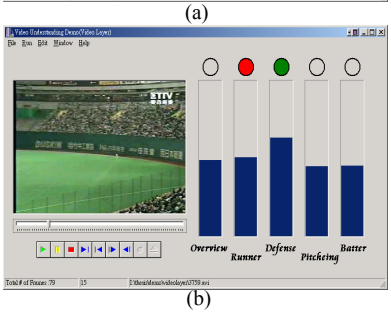
Table 2-2. Experimental result for category level.

Video Class	Detection Accuracy	False Alarm
Overview	96.23%	26.79%
Runner	97.20%	41.49%
Defense	96.55%	28.07%
Pitching	98.70%	38.92%
Batter	98.70%	43.60%

Figure 8a shows an example for the identified semantic features of the pitching scene. Figure 8b illustrates the video understanding results for the scene of defense. The performance of our system has been shown in Tables 1~2.



(a)



(b)

Figure 8. (a) The semantic interpreter, (b) The category classifier.

7. CONCLUSIONS

The main contribution of this paper is to model an inference system by constructing the relationship between unobservable concepts and observable concepts. We use the visual significance of an object as a prior and the condition probabilities to determine the parameters of the BBNs. Given some low level evidences to the BBN, which consists of links of nodes, the system will generate high-level semantic meaning of the video content.

REFERENCE

[1] H. Zhang, A Wang, and Y. Altunbasak, " Content-based video retrieval and compression, " IEEE ICIP, Oct. 1997.

[2] D. Zhong and S. F. Chang, "Spatio-temporal video search using the object-based video representation," IEEE ICIP, Oct. 1997.

[3] M. R. Naphade, I. Kozintsev, and T. S. Huang, "A Factor Graph frame work for demantic indexing," IEEE Trans. on CAS for VT, pp.40-52, Jan. 2002.

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

[5] F. V. Jensen, K. G. Olesen, and S. K. Andersen, "An Algebra of Bayesian Belief Universes for Knowledge-Based Systems," Networks, vol. 20, pp.637-659, 1990.

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

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

[8] 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.

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

[10] M. K. Kim, E. Kim, D. Shim, S. J, G. Kim, "An Efficient Global Motion Characterization Methods for Image Processing Application", IEEE Trans, on Consumer Electronics, Vol. 43, No. 4, Nov. 1997.

[11] B. S. Manjunath, J. R. Ohm, V. V. Vasudevan, A. Yamada, "Color and Texture Descriptors", IEEE Trans. On CAS for VT, Vol. 11, No.6, June 2001.