# Image Analysis and Interpretation for Semantics Categorization in 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 semantic information in videos is useful for content-based video retrieval and summarization. Traditional image/video understanding is formulated in terms of low-level features describing the structure and intensity of the input image/video. How to generate the high-level knowledge such as common sense and human perceptual knowledge is one of the most difficult problems This paper attempts to bridge this gap through the integration of image analysis algorithms and multi-level Semantic Network (SN) to interpret the semantic meaning of the baseball video.

**Keywords:** Image/Video understanding, Bayesian network, Multi-level semantic network.

## 1. Introduction

Recently, a large amount of digital media including image, audio, video, streaming video clips, panorama images and 3D graphics have been produced daily. 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] proposed a novel probabilistic framework (multijects and multinets) for semantic indexing and retrieval in digital video. They apply the factor graphical model to provide an efficient framework for video inference.

The pioneer work of Vasconcelos and Lippman [4] introduces 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]. Domain-specific knowledge is characterized by the network structure in terms of the causal/relevant relationships between multiple variables, which are represented by the complicated joint probability distributions that can be simplified as a set of conditionally independent relationships.

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 Semantic Network (SN) 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 SN will infer the high-level semantic of the video. In sports, the events that unfolded are governed by the rules. 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.

We need to interpret and then encode the video in high-level knowledge such as common sense and human perceptual knowledge. The high-level knowledge is represented in a form of abstraction or summarization, a non-geometric representation. Different from the previous researches, this paper proposes a multi-level semantic 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. 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 SN may infer certain high-level concepts.

We develop several SNs to model the different semantic events in the baseball video (see Figure 1).

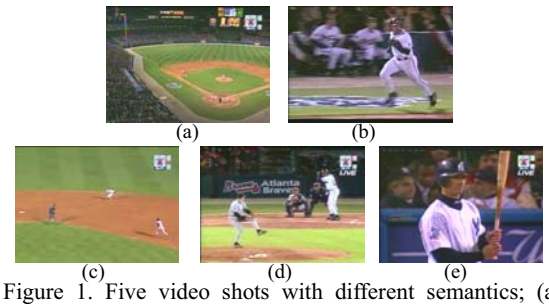


Figure 1. Five video shots with different semantics; (a) overview, (b) runner snapshot, (c) defending view, (d) pitching view; (e) batter snapshot.

Having applied the BBN training procedure on the SNs, we may 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 interpret the input video with high-level semantic meaning. Specific domains contain rich spatial and temporal transitional structures. 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*.

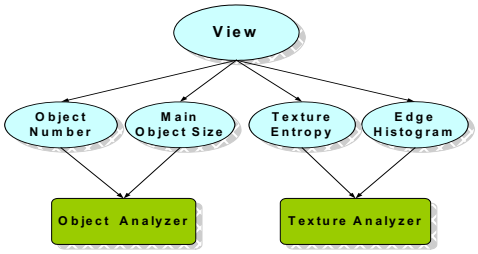



Figure 2. A SN for the semantic concept *"View"*.

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 2, 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 t*exture-analyzer*. These low-level evidences are connected to some image analyzers, such as t*exture-analyzer*, o*bject-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.

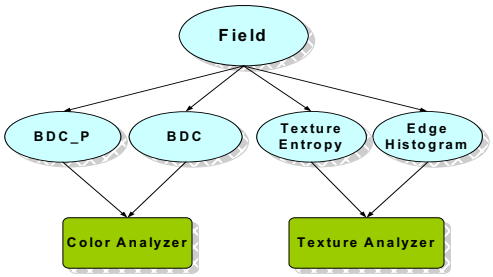



Figure 3. A SN for the semantic concept *"Field"*.

Figure 3 illustrates another SN that may be used to infer the mid-level semantic *Field*. The t*exture-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 c*olor analyzer* provides 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.

The mid-level semantic concept related to motion activity is described by the SN shown in Figure 4. The displacement vector information is analyzed so that it can be used for SN to demonstrate the certainty of three different mid-level semantics: fast panning, regular panning and zooming.

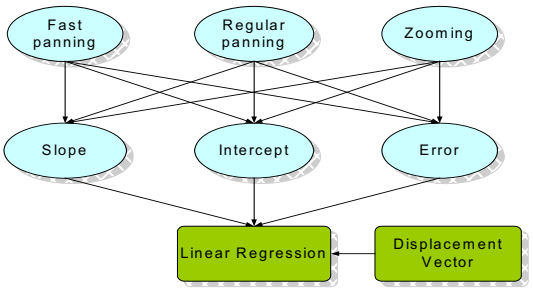



Figure 4. A SN for the semantics concepts "*fast-panning*", "*regular-panning*" and "*zooming*".

On the top of the multi-level BBN hierarchy are the root nodes representing the certainty of the five different categories: *Overview, Runner, Defense, Pitching* and *Batter*. The upper level of the multi-level network is shown in Figure 5 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.

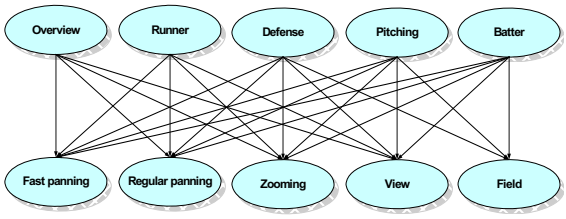



Figure 5. A SN for the mid-level semantics.

Here, we use five mid-level nodes to represent semantic concepts such as *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 6 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.

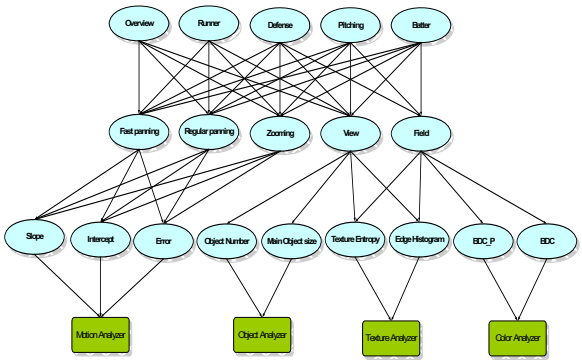



Figure 6. An overall SN for baseball video.

## 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. For video summarization (or video understanding), the first step is to segment the video into temporal "shot". The second step is to extract the low-level evidences, which consist of the object motion, colors, textures and the panning motion of the camera.

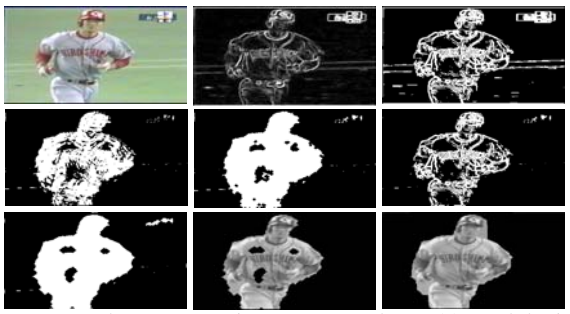
**1) Scene Change Detection.** First, the input video is segmented into video shots, which represent a sequence of frames captured from a unique and continuous record from a camera. Adjacent frames of the same shot exhibit temporal continuity. The main problem of segmenting a video sequence into shots is how to distinguish between scene breaks and normal changes that happen in the scene.

**2) Texture.** The gray-level co-occurrence matrix (GLCM) is used to capture the spatial dependence of gray-level region. The co-occurrence matrix method of texture description is based on the repeated occurrence of some gray-level configuration in texture. 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]. The edge information is categorized into five classes: vertical, horizontal, $45^{\circ}$ diagonal, $135^{\circ}$ diagonal, and isotropic.

**3) Color.** 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.

**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. 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.

**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 video analysis, and most of the edge information is redundant. We assume that the objects are moving, and the background scene is complex but stationary. The step-by-step results of the segmentation processes are illustrated in Figure 7.



| | | |
|---|---|---|
| a | b | c |
| d | e | f |
| g | h | i |

Figure 7 The segmentation (a) original video;(b)edge image; (c)after thresholding; (d) after accumulation; (e) after closing operation ; (f) result of c AND e; (g) after closing on (f); (h) after noise removal; (i) after region growing.

## 4. Bayesian Belief Network

To train the SNs, we apply the techniques in Bayesian Belief Network (BBN) which has been proved to be an effective statistical model for knowledge representation and inference. SN is similar to BBN which 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.

The knowledge of the domain is used to construct the network. One of the major problems is how to treat the uncertainty in the specific knowledge domain. BBN process has been proved to be a powerful mechanism to model the uncertainty and the reasoning in terms of some quantity measurements. In BBN, direct arcs between variables represent conditional dependency. 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 $\mathbf{X}=\{x_1, x_2,..., x_n\}$, 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 $\mathbf{Pa}_i$ 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

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

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

Inference, 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.

## 5. 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 which indicates different semantic content. Moreover, each shot may consist of different number of image sequences. We have collected many video shots for each of the five different semantic meaning. Then, we divide the video shots of the same semantic into two groups, one for training and the other for testing. The size of each color image frame is 352 × 240, and its frame rate is 30 frames per second. Each shot lasts about 1 to 2.5 seconds.

In the experiments, we assume that all video sequences can be partitioned by the scene change detection. We use UPMOST mpeg-1 encode card to grab the fast-moving object without causing too much noise and blurring effects on the videos of the TV programs. 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.

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 8 shows an example for the identified semantic features of the pitching scene. Figure 9 illustrates the video understanding results for the scene of defense. The performance of our system has been shown in Tables 1~2. There are five video categories in our experiments, and the system has some false alarms mentioned in the following:

(1) *Defense* scenes are misinterpreted as the case of *runner* and *overview* because both of them are sub-network of *defense*. The false alarm of *runner* video occurs if the *zooming* semantic is not correct or *Field* is found as outfield. The false alarm of *overview* shot happens when *zooming* and *fast panning* feature is not accurate.

(2) *Batter* snapshots are misidentified as the cases of *pitching* and *runner*. The crucial semantic to discriminate *batter* from *pitching* is *view*. If the pitcher is too closed to the batter and the catcher, the false alarm occurs. The false alarm of *runner* shot occurs when *fast panning* and *regular panning* are misidentified.

(3) *Pitching* scenes are misidentified as the *overview* shots because *overview* is a sub-net of *pitching*. The false alarm occurs when the *regular panning* is not accurate or the *field* is mistreated as an infield.

(4) Overview scenes are often misinterpreted as *runner* snapshots. They have the same semantic variables. False alarm occurs if *view* and *fast panning* is simultaneously wrong.

(5) Runner snapshots are misidentified as *overview*, *defense* and *batter*. The reason for the false alarm of *overview* and *defense* is similar to the above reasons. The false alarm of *batter* occurs, when the batter in close-up view performs fast actions such as waving his bat, which will be mistreated as a panning motion.

## 6. Conclusions

The main contribution of this paper is to propose an inference system by constructing the SNs which are based on 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 SN. Given some low level evidences to the SN, which consists of links of nodes, the system will generate high-level semantic meaning of the video content.

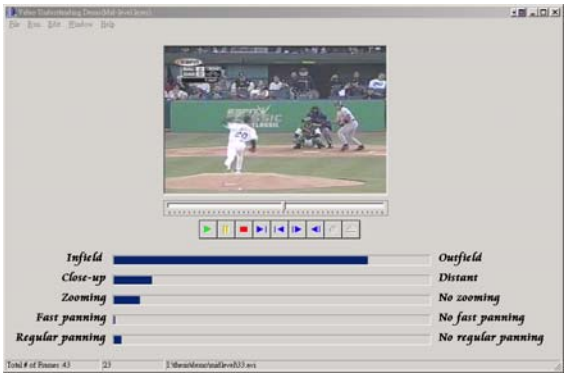



Figure 8. The semantic interpreter.

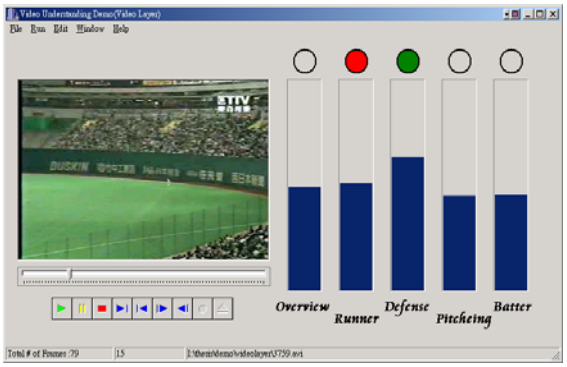



Figure 9. The category classifier.

## 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 Circuits and System for Video Tech., Vol. 11, No.6, June 2001.

[12] C. L. Huang, S. H. Jeng, "A model-based hand gesture recognition system", Machine Vision and Applications, Vol 12, Issue 5, 2001, pp 243-258.

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