From Motion Patterns to Visual Concepts for Event Analysis in Dynamic Scenes

Lun Xin and Tieniu Tan

National Laboratory of Pattern Recognition,
Institute of Automation, Chinese Academy of Sciences, 100080 Beijing, P.R. China
{lxin, tnt}@nlpr.ia.ac.cn

Abstract. The analysis of events in dynamic scenes has become an important and challenging problem increasingly in recent years. Events can be considered as obvious changes of important features with semantic meanings. From this viewpoint, the fundamental task of events analysis is to extract semantically meaningful changes and associate all of these basic motion patterns and changes with relevant visual concepts of moving objects in dynamic scenes. In this paper, we propose a method to extract lower level motion patterns and associate them with visual concepts respectively in a well-defined structure. Furthermore we also analyze latent spatial-temporal relationships among these basic visual concepts for event modeling and analysis. Finally, we present experimental results which prove the effectiveness of our approach on some real-world videos of dynamic scenes.

1 Introduction

As a challenging problem, semantic analysis of dynamic scenes has been paid more attention by researchers in recent years. Furthermore many methods have been presented for dealing with it. Some of these methods define and analyze semantic meanings based on the global statistical properties of the movement. From the global viewpoint, this kind of methods usually ignores semantics of features exhibited in a lesser temporal scale. On the other hand, considering basic semantic meaningful features in small temporal interval is useful for the semantic understanding of the entire event. The basic flowchart of a video surveillance system will include elementary procedures such as environment modeling, object detection, tracking and recognition. However, each of these is not the termination of semantic analysis in a dynamic scene; there should be some further missions for achieving semantic understanding and interpretation of what behaviors or events performed by those moving objects in this dynamic scene. Compared with the lower level processing, the higher level phase involves spatio-temporal relationship mining, reasoning under uncertainty, semantic representation, and so on [1].

The basic requirement of the event analysis is to extract semantically meaningful motion patterns in the scene [2]. In different research areas, semantics has quite different meanings. There is a restrictive definition in semiotics that semantics implies the relationship between signs and objects. But for language science, semantics means the meaning and relationship of words. In our research work, we adopt the definition that semantics is the mapping and integration between related concepts [3].

P.J. Narayanan et al. (Eds.): ACCV 2006, LNCS 3851, pp. 826 – 835, 2006.

But there is a gap between measurable features and semantic meanings. According to the ability and the procedure of human in perception and understanding for the world, event can be considered as the semantically meaningful changes in the scenes. The basic elements for event analysis and understanding are various concepts. Each concept denotes a special semantic meaning. And all these concepts are grouped into different clusters according to their semantic functions. For the purpose of semantic analysis and understanding of events in dynamic scenes, all related concepts should be obtained firstly, and all these concepts should be organized in a well-defined structure.

As declared by some genres in philosophy, the world can be considered as the integration of different kinds of entities. From this viewpoint, all existing things in a special dynamic scene, such as different regions, moving or static objects can be treated as different entities with their own relative properties. Further more, given concepts can be used to denote these entities and their properties. The semantic analysis in the special domain can be achieved from these concepts and their relationships.

The three fundamental components of a concept are an entity, a term or a word and corresponding attributes [4]. Each concept is described as a sign by a term or a word to distinguish each other. And the difference or the similarity of different concepts can be defined on all these measurable attributes.

The difficulty for a certain definition of event is due to various demands from different domains. Thibadeau [5] defines first-order change descriptions as motion and the second-order ones as action, and Newtson [6] treats activity as the maintenance of first-order primitive properties. In this paper we consider events as obvious changes of important features as mentioned in [7].

High-level analysis and understanding of dynamic scenes is the final goal of computer vision. Compared with the traditional vision tasks such as tracking and recognizing moving objects, high-level vision is to achieve deeper analysis of spatial-temporal relationships exhibited by all visible and measurable data in dynamic scenes [8]. Contextual spatial-temporal information acts as an important clue for semantic understanding.

This paper proposes a method to associate semantic meaningful motion patterns with corresponding visual concepts for semantic analysis of events in dynamic scenes. Sections in this paper are organized as follows. Section 2 outlines previous work of event modeling and analysis. Methods for motion pattern extraction and concept modeling are described in Section 3 and Section 4 respectively. Then experimental results are showed and analyzed in Section 5. Finally, we draw conclusions and discuss future work in Section 6.

2 Previous Work

Existing work on event analysis is usually based on trajectory analysis of moving objects. Methods for trajectory extraction and simple object classification are based on some traditional methods proposed in [10, 11, 12, and 13]. Since more expressive semantically meaningful features can be extracted from trajectories, they are not organized in a proper structure for farther semantic analysis. That means each semantically meaningful feature should be associated with a concept, and the relationship of these concepts should also be considered seriously.

In [14], events are modeled and recognized by exhibited periodically variational patterns. Similar work proposed in [15] treats human activities as descriptions of their

basic spatial-temporal characteristics. Ivanov and Bobick [16] extract primitive features by using HMM and recognize activities with a context-free parsing mechanism. Event or activity can also be divided into elementary components, and can be detected, represented and identified at different levels in a uniform framework [17, 18, 19, 20, and 21]. Kojima et al. [21] employ a case frame with syntactic components to model events in office scene. All syntactic components are associated with related semantic features, and the model can provide natural language descriptions of those official events. Chaudron et al. [22] represent the interpretation of event in dynamic scene as a symbolic layered prototype by Petri nets.

In recent years, more and more researchers tend to use probabilistic frameworks to express and analyze events, such as Bayesian networks, hidden Markov models, etc. All these models have a common peculiarity that stochastic parameters can be acquired automatically without any assumptions of prior knowledge under uncertainty. Considering idiographic demands under different circumstances, some variations have emerged. Galata et al. [23] mention a method to present human behavior by variable length Markov models (VLMM). The algorithm of coupled hidden Markov models (CHMM) to model two-handed interactions is presented in [24]. At the same time, the superiority of these methods mentioned above brings obvious shortages. The computation of parameters for the given structure of a model is time-costly. To fit another problem, the structure of the model must be changed, and the learning for variable structures is more difficult.

From Birnbaum et al., who use ontology to define causal changes in their attention controller in [25] , ontology related methods [26, 27, and 28] are increasingly applied in various areas, such as semantic web, data mining, knowledge management, information fusion, linguistics and etc.

3 Motion Patterns Extraction

In a visual surveillance system, scenes of the environment captured by fixed cameras can be looked as combinations of all kinds of visual entities exhibited in the video data. These entities are regions with different spatial positions and appearances, moving objects and their different motion and interaction patterns, and so on. The semantic analysis of the scene can be looked as mining and analysis for all kinds of relationship of related visual concepts. So at the beginning of this kind of work, all visual concepts must be defined and constructed in a unified form.

3.1 Location States Extraction

To determine locations of moving objects in a dynamic scene, we can look the scene as different adjacent regions which are labeled with their attributes, such as grassplot, road, sideway, intersection, crosswalk, etc. There is a pre-hypothesis that the semantic attribute of each region blob in the scene is homogeneous. So each region has its unique semantic label. Examples of semantic attributes of different labeled regions are showed in Table 1. At the same time, different spatial regions have invariable topological adjacent relationship under a fixed camera. Figure 1 shows an example of topological adjacent relationship for different regions in a special scene.

Table 1. Semantic Attributes of Labeled Regions

Labeled Regions	Semantic Attributes
Road	Vehicles and other moving objects can move in it.
Sideway	Only allow foot passengers moving in it.
Grassplot	Any motion of moving objects occurs in it is not allowed.
Crossway	Any stop of moving objects in it is not allowed.
Parking Lot	Only allow vehicle parking in it.

As illustrated in Figure 1, nine nodes denote nine different regions, and edges refer the adjacent spatial relationship of these regions. Different color of these nodes means different semantic attributes of these regions and highlighted edges indicate that moving objects can transit between two connected nodes. All related constraints can be defined in this topological graph.

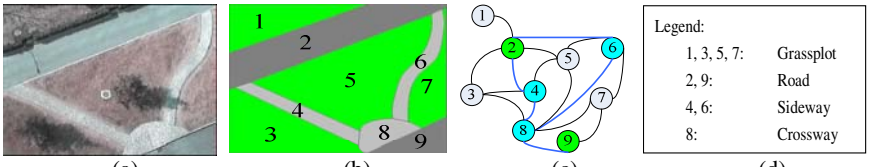

Fig. 1. Topological Relationship of Regions in the Scene

We use central points of moving regions as the approximate locations of moving objects. Mapping coordinates of central points to the semantically labeled image, we can obtain regions objects occupied. When objects move through different regions, label sequences of region transitions can also be obtained.

3.2 Motion States Extraction

All moving objects are leading actors in dynamic scenes. Event modeling and semantic analysis are focused on them. We can extract and express motion states of moving objects separately. Under a fixed camera, a trajectory of a moving object is represented as temporally sequential pairs of coordinates in frames. These pairs of coordinates can be presented like this format:

$$L = \{(x_1, y_1), (x_2, y_2), \cdots, (x_t, y_t), \cdots\} \tag{1}$$

where (x_t, y_t) is the coordinates of a moving object at time t or is at the t th sequence number of the current frame.

The basic motion states of a single moving object are "*Move*" and "*Stay*", and the basic direction states are "*Go Straight*", "*Turn Left*" and "*Turn Right*". The small trajectory segments of moving objects with the temporal scale about two seconds (50~60 frames) can be divided into these basic elements.

Figure 2 shows an example of two moving objects separately, labeled by m and n. When each moving object has appeared in the scene, a sub-coordinate is set up for

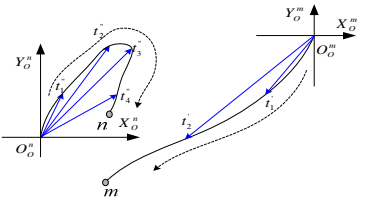

Fig. 2. Motions in Different Coordinates

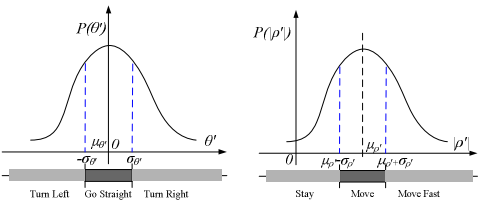

Fig. 3. Motion States Mapping

this object, and all related motion status can be extracted and calculated in this sub-coordinate. The origin of each sub-coordinate is the initial position of each moving object. For easy calculation, we present a trajectory in Polar Coordinate,

$$L_{\rho-\theta}=\{(\rho_1,\theta_1),(\rho_2,\theta_2),\cdots,(\rho_t,\theta_t),\cdots\} \tag{2}$$

and define the interval change ρ' and θ' as

$$\begin{aligned}\rho'&=\{(\rho_{i+l}-\rho_i)\},\ \rho'\sim N(\mu_{\rho'},\sigma_{\rho'}{}^2)\\ \theta'&=\{(\theta_{i+l}-\theta_i)\},\ \theta'\sim N(\mu_{\theta'},\sigma_{\theta'}{}^2)\end{aligned} \tag{3}$$

where l is an interval.

Based on statistical analysis of training data, we make an assumption that $|\rho'|$ and θ' obey the Gaussian distribution under the existing noise. The motion patterns about movement M_{Status} and direction Dir_{Status} can be mapped into different status as shows in Figure 3. All these parameters are all learned from videos of special scenes under special viewpoints. As a result, when zoom ratio or viewpoint changed, all these parameters should also be recalculated.

3.3 Interaction States Extraction

When we analyze the interaction between moving objects, we should consider opposite distances of these objects in a unique coordinate of the whole image (see object *i*, *j* and their opposite distance in Figure 4. (a)). The basic varieties of opposite distances can be increase, reduce and without obvious changes. By using learned thresholds, we can distinguish opposite distance *d(i,j)* between object *i* and *j* as one of those three basic varieties. And all these thresholds are also view or scale based. When several moving objects are very close to each other, it is hard for our tracking algorithm, even for human, to determine whether they should belong to a whole moving object or regard as separate objects. In the same way, when objects are so far from each other, it is unnecessary for considering their interactions. To deal with this problem, we can define different region scales (Figure 4. (b)) for each moving object. The size of each region scale is related to the size of each moving object. By using these region scales, we can determine interaction states of moving objects easily.

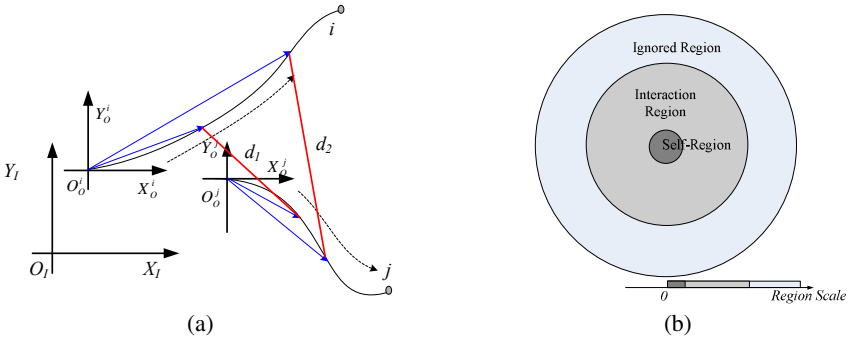

Fig. 4. (a) Opposite Distances between Objects. (b) Region Scale for Interactions.

4 States Transition-Based Concept Modeling

The definition of concept in Webster's dictionary is "an abstract or generic idea generalized from particular instances", it is the basic element of human thought. As a symbolic abstraction of the essence of reality, a concept contains some related measurable attributes. It is exhilarative that location states, motion states and different interaction categories of moving objects mentioned above are all based on measurable attributes. For further semantic analysis of events performed by moving objects in dynamic scenes, we should associate all these states, patterns and categories with corresponding concepts in certain temporal sequences. Some of concepts and verbs used in our model are chosen from the classification of motion verbs in traffic scene given by Badler [9] formerly.

All related visual concepts can be defined on transitions among those states. Figure 5. illustrates transitions on the basic states, such as "*Move*", "*Stay*", "*Go Straight*", "*Turn Left*" and "*Turn Right*". These transitions can present all semantically meaningful features of moving objects. In each temporal scale, motion patterns can be classified into basic semantic states, and we can obtain temporal sequences of those states showed below, and some related visual concepts can be associated with different segments of states transition sequences.

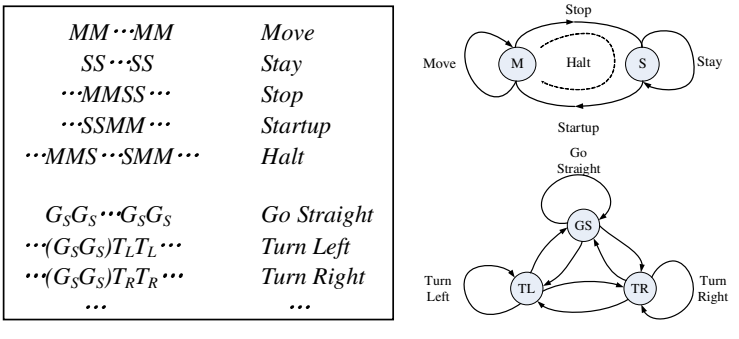

Fig. 5. Transitions Model of Basic States

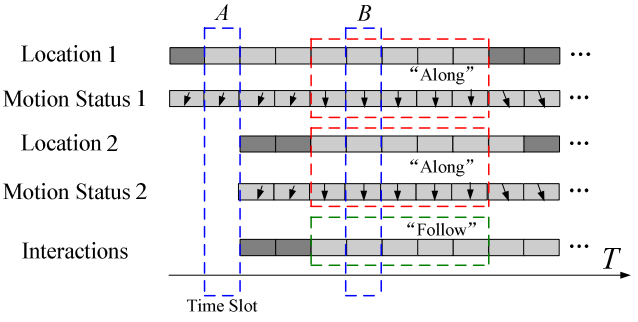

Fig. 6. Conceptual Vectors and Semantic Representations

By using this method, we can obtain all kinds of related visual concepts. For the aspect of motion, we can describe it as "*Go Straight*", "*Turn Right*", "*Turn Left*", "*Retrace*", etc. And there are different interactions in the scene. Interactions between moving objects and special regions can be represented as spatial relationships, such as "*Occupy*", "*Enter*", "*Transfer*", "*Appear*", etc. Interactions of two moving objects can be "*Close To*", "*Away From*", "*Encounter*", "*Follow*", "*Retrace*", etc. And we should choose different concept for vehicles and passengers.

In a certain time slot, we can integrate all these obtained visual concepts into a conceptual vector (see Figure 6.). In this figure, each block denotes a corresponding visual concept with different color, and each arrow expresses different motion direction of this moving object in a certain time slot. By using each conceptual vector, simple semantic representation of event performed by moving objects can be obtained.

All concepts in visual surveillance are obtained at different scales. That means some basic concepts are components of other concepts, such as "*Move*" and "*Halt*", "*Go Straight*" and "*Retrace*". So concepts with similar meanings can be presented in dendriform structure as different clusters.

5 Experimental Results

From our multi-camera visual surveillance system, we choose two fixed cameras which can capture wide visual fields from taller points of views. Under this condition, the influence of 3D to 2D perspective can be reduced, and then we can use the coordinates of moving objects in the image plane as the probable positions of them in the real scene.

According to the method mentioned above, we calculate all parameters from training video, and then analyze all related motion patterns of moving objects in the certain temporal scale. After associating these motion patterns with corresponding visual concepts in conceptual vectors, simple semantic representations will be obtained by using these related concepts in a time slot.

Figure 7 explains complex events performed by moving objects in two selected scenes and shows simple semantic representations of these events. As showed in Figure 6., we will select "Along" to express the motion of an object if it moves unremittingly in the same region without obvious direction change in several adjacent time slots. In the same way, "Follow" will be adopted if two objects are moving in the similar direction and their opposite distance keeps reducing.

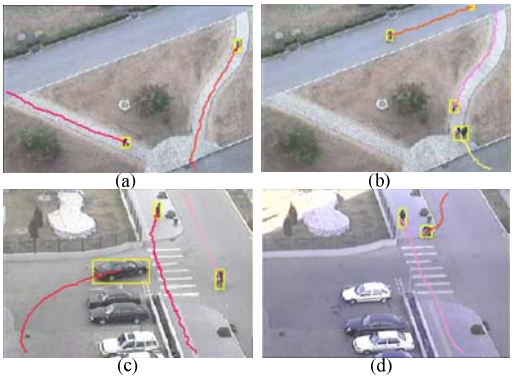

Fig. 7. Complex Events and Simple Semantic Representations

6 Conclusion and Future Work

In this paper, we have presented a method to associate motion patterns with corresponding visual concepts for event analysis seen in dynamic scenes. The key points of our method are extraction of motion patterns, concept generation and modeling. Simple semantic representations of events in the dynamic scenes are obtained in some real world videos, and the result also validates the effectiveness of this method.

Manually labeling of different regions in the dynamic scene and negligence the uncertainty of observed data are main limitations of our methods. In the future, we will adopt some learning methods to achieve semantic labels by using texture and motion information. To handle the uncertainty problem of our method, probabilistic mechanism should be a good choice. Extended experiments and embedded polishing are also needed for our method.

Acknowledgement

The work reported in this paper was funded by research grants from the National Basic Research Program of China (No. 2004CB318100), the National Natural Science Foundation of China (No. 60335010) and the International Cooperation Program of Ministry of Science and Technology of China (No. 2004DFA06900).

References

[1] Mubarak Shah: Guest Introduction: The Changing Shape of Computer Vision in the Twenty-First Century. International Journal of Computer Vision. Vol. 50. No. 2. (2002) 103-110

[2] A. Ekinci, A. M. Tekalp: Generic Event Detection in Sports Video using Cinematic Features. In Second IEEE Workshop on Event Mining (EVENT'03). (June 2003) 17-24

[3] Fauconnier, G.: Mapping in Thought and Language. Cambridge University Press. (1997)

[4] Dahlberg, I.: Conceptual Definitions for Interconcept. International Classification. Vol. 8. No. 1. (1981) 16-22

[5] R. Thibadeau: Artificial Perception of Actions. Cognitive Science. 10(2). (1986) 117-149
[6] D. Newtson: Foundations of Attribution: the Perception of Ongoing Behaviour. New Directions in Attribution Research. Laurence Erlbaum. Hillsdala. NJ. (1976) 147-223
[7] R.J. Howarth, H. Buxton: Conceptual Descriptions from Monitoring and Watching Image Sequences. Image and Vision Computing. Vol. 18. (2000) 105-135
[8] Bernd Neumann: A Conceptual Framework for High-level Vision. Bericht. FB Informatik. FBI-HH-B245/02. (Juli 2002)
[9] Badler, N.I.: Temporal Scene Analysis: Conceptual Descriptions of Object Movements. Technical Report No. 80. Dept. of Computer Science. University of Toronto. (1975)
[10] S.S. Intille, J.W. Davis, A.F. Bobick: Real Time Closed World Tracking. IEEE Proc. Computer Vision and Pattern Recognition. (1997) 697-703
[11] A. J. Lipton, H. Fujiyoshi, R.S. Patil: Moving Target Classification and Tracking from Real Time Video. Proc. Fourth IEEE Workshop Application of Computer Vision. (1998) 8-14
[12] I.Haritaoglu, D.Harwood, L.S.Davis: W4: real-time surveillance of people and their activities. IEEE Transactions on Pattern Analysis and Machine Intelligence. Vol. 22. Issue: 8, (Aug. 2000) 809-830
[13] C.R. Wren, A. Azarbayejani, et al.: Pfinder: Real Time Tracking of the Human Body. IEEE Trans. Pattern Analysis and Machine Intelligence. Vol. 19. No. 7. (July 1997)
[14] L. Davis, R. Chelappa, A. Rosenfeld, D. Harwood, I. Haritaoglu, R. Cutler: Visual Surveillance and Monitoring. In DARPA Image Understanding Workshop. (1998) 73-76
[15] A. Galton: Towards an Integrated Logic of Space, Time and Motion. Proc. International Joint Conf. Artificial Intelligence (IJCAI). (Aug. 1993)
[16] Yuri A. Ivanov, Aaron F. Bobick: Recognition of Visual Activities and Interactions by Stochastic Parsing. IEEE Trans. on Pattern Analysis and Machine Intelligence. Vol. 22. No. 8. (Aug. 2000) 852-872
[17] Dance, S., Caelli, T.: A Symbolic Object-oriented Picture Interpretation Network: SOOPIN. In Advances in Structural and Syntactic Pattern Recognition. Proceedings of the International Workshop. H. Bunke, Ed. World Scientific Publishing Co. (1993) 530-541
[18] M. Haag, H.-H. Nagel: Incremental Recognition of Traffic Situations from Video Image Sequences. Image and Vision Computing. Vol. 18. (2000) 137-153
[19] R. J. Howarth, H. Buxton: Conceptual descriptions from Monitoring and Watching Image Sequences. Image and Vision Computing. Vol. 18. (2000) 105-135
[20] Nuria Oliver, Ashutosh Garg, Eric Horvitz: Layered Representations for Learning and Inferring Office Activity from Multiple Sensory Channels. Computer Vision and Image Understanding. Special Issue on Event Detection in Video. Vol. 96. Issue 2. (Nov. 2004) 163-180
[21] Atsuhiro Kojima, Takeshi Tamura, Kunio Fukunaga: Natural Language Description of Human Activities from Video Images Based on Concept Hierarchy of Actions. International Journal of Computer Vision. Vol. 50. No. 2. (Nov. 2002) 171-184
[22] Laurent Chaudron, Corine Cossart, Nicolas Maille, Catherine Tessier: A Purely Symbolic Model for Dynamic Scene Interpretation. International Journal on Artificial Intelligence Tools. Vol. 6. No. 4. (Dec. 1997) 635-664
[23] Aphrodite Galata, Neil Johnson, David Hogg: Learning Structured Behaviour Models Using Variable Length Markov Models. Computer Vision and Image Understanding (CVIU) Journal. Vol. 81. No. 3. (March 2001) 398-413
[24] M. Brand, N. Oliver, A. Pentland: Coupled Hidden Markov Models for Complex Action Recognition. Proc. of the Conference on Computer Vision and Pattern Recognition (CVPR '97). (1997) 994-998

[25] L. Birnbaum, M. Brand, P. Cooper: Looking for Trouble: Using Causal Semantics. Proceedings of the Fourth International Conference on Computer Vision. Berlin. Germany. IEEE Computer Society Press. Silver Spring. MD. (1993) 49-56

[26] Christopher Town: Ontology-driven Bayesian Networks for Dynamic Scene Understanding. Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops (CVPRW'04). (27-02 June 2004) 116-123

[27] P. Varga, T. Mészáros, Cs. Dezsényi, T.P. Dobrowiecki: An Ontology-based Information Retrieval System. The 16th International Conference on Industrial & Engineering Applications of Artificial Intelligence and Expert Systems. Loughborough. U.K. (23-26 June 2003)

[28] Yanmei Wang, Zhonghua Yang, Pe Hin Hinny Kong, Robert Kheng Leng Gay: Ontology-based Web knowledge management. Proceedings of the 2003 Joint Conference of the Fourth International Conference on Information, Communications and Signal Processing. Vol. 3. (15-18 Dec. 2003) 1859 -1863