http://ansinet.com/itj

ITJ

ISSN 1812-5638

INFORMATION TECHNOLOGY JOURNAL

ANSInet
Asian Network for Scientific Information
308 Lasani Town, Sargodha Road, Faisalabad - Pakistan

Information Technology Journal 12 (1): 113-119, 2013
ISSN 1812-5638 / DOI: 10.3923/itj.2013.113.119

Semantic Event Mining Based on Hierarchical Structure for Soccer Video

[1]Pixi Zhao and [2]Xiukun Wang
[1]School of Computer Science and Engineering, Dalian Nationalities University, Dalian 116600, China
[2]School of Computer Science and Technology, Dalian University of Technology, Dalian 116024, China

Abstract: A three-layer framework of Semantic event Mining for soccer video was proposed, namely shot semantic layer, scene semantic layer and semantic event layer from bottom to top. The function of shot semantic layer is the classification of physical shots. As a result, the streaming video data composed of original physical shots are translated into a character sequence, in which each character means a kind of shot type, such as global shot (A), player's gathering shot (G), a player shot (P), a referee shot (C), spectator shot (S), replay shot (R) and other shot (O). At the scene semantic layer, the character sequence is segmented into scene subsequences. Unlike most existing scene segmentation approaches by shot-clustering, a novel method was also proposed in this study which uses global shot as the boundary of a scene for scene segmentation. The task of semantic event layer is the recognition of a semantic event corresponding to a scene subsequence. First, GSP algorithm is used to mine out the standard sequence mode for each known semantic event from the training data set. Second, the frequent mode of a scene subsequence waiting for recognition is mined out by APRIORI algorithm. Finally, the sequence distance between the standard sequence mode of a known semantic event and the frequent mode of a scene subsequence waiting for recognition is computed and semantic event recognition for this scene subsequence is accomplished based on the distance. Experiment results aimed at mining for goal event and shooting event in soccer video have shown that the proposed framework is feasible and effective.

Key words: Three-layer framework, scene segmentation, semantic event mining, generalized sequential patterns, sequence distance

INTRODUCTION

Semantic understanding of videos is quite a challenging problem due to the semantic gap. It is not feasible to expect a general solution that will work successfully across all genres of video, much of the prior art concerns genre specific approaches (Jin *et al.*, 2004). Semantic events of video, normally defined as the interesting events ,capture user attentions, e.g., a soccer goal event. A number of recent research works have been focused on news and sports videos for event detection by data mining techniques (Thirumurugan and Hussain, 2012). Occurring events in news videos usually implicate some semantic information. For example, a caption event will greatly help people to understand the news contents, the close-up picture of President Hu Jingtao and president Barack Obama implies a news about relations between China and the United States. A event mining method for medicine videos was proposed by Zhu *et al.* (2003). Through video structure analysis three kinds of event (dialogue, presentation, clinical operation) have been mined. A method for mining association rule between shots in news videos was proposed by Zhu and Wu (2003). The classification label of each physical shot was obtained by shot clustering after the segmentation of physical shots, so the original physical shot sequence was translated into a symbol sequence. Then the frequent mode of the symbol sequence was mined out by APRIORI algorithm which is the main content of the news video, e.g., the sequence of a anchor shot by a guest shot. This unsupervised association mining method corresponds to novelty and has lower time complexity compared with other approaches based on statistical model. However, it can not be directly applied to sports video, because there are not story of the obvious plots similar to news video in sports video. Furthermore, the events of sports video are mostly rare events, e.g., goal event in soccer video.

With the great number of audiences of sports video and rapidly increasing sports videos on different media, the sports video analysis has become an important problem owing to its great potentially commercial value. For example, extracting highlights in soccer video to make it possible to deliver video clips into mobile devices or Website. Moreover, summarizing a long sport video into short segments will bring convenience to content-based

Corresponding Author: Pixi Zhao, Graduate Student Office, Dalian nationalities University, Dalian Development Zone, Liaoning Province, China, 116600 Tel: 0086-13009430833 Fax: 0086-0411-87630449

video retrieval and then to facilitate the users finding the preferred video content in a database. These inspire the researches on sport video summarization and highlights detection. However, detecting events in general sports video is still an open problem at present, most of the researcher focus on a special kind of sports video, including American football, basketball, etc. in which soccer video is mostly concerned for its high audience rating. In the case of soccer sports video, a rule based algorithm using low-level audio/video features for soccer video summarization was proposed by Li *et al.* (2003). In (Lefevre *et al.*, 2002), audio data were divided into short sequences which are classified them into three classes such as speaker, crowd, referee whistle. Event detection by recognizing the textual overlays from soccer video was proposed by Babaguchi *et al.* (2004). Dominant speech features were used to generate soccer highlights by Wan and Xu (2004). Ding and Fan (2007) proposed segmental Hidden Markov Mode for view-based soccer analysis was proposed. HMM based framework to fuse audio and video features to recognize the play and break scenes in soccer video sequences was proposed by Barnard *et al.* (2003). A HMM based framework to extract 'attack' scene from football video has been proposed by Ren and Jose (2005). Support Vector Machine based event detection for soccer video was proposed by Ancona *et al.* (2001). Finite State Machine based annotation of soccer video was proposed by Assfalg *et al.* (2004). Semantic notion of offense for event detection of soccer video was proposed by Wang *et al.* (2004). The technique to retrieve the football video clips by using its global motion information was proposed by Yu and Zhang (2001). Recently, researchers have presented the use of tracking the positions of players or ball for soccer video analysis (Xu *et al.*, 2004; Li *et al.*, 2005; Nillius *et al.*, 2006). There have been many successful works in soccer video analysis as mentioned above. The approaches based on statistical models like HMM are usually time-consuming and also faced with the training problem of rare events. Although, SVM presents promising generalization performance, its training process does not scale well as the size of the training data increases. The rare event detection issue in soccer videos was addressed through the proposed data mining framework via the combination of distance-based and rule-based data mining techniques (Thirumurugan and Hussain, 2012). Seventeen low-level features were directly used for the detection of higher semantic events, so that there is certain shortage in bridging the semantic gap in this framework. A novel hierarchical framework for soccer video event sequence detection and classification was proposed by Kolekar *et al.* (2009). Unlike most existing video classification approaches which focus on shot detection followed by shot-clustering for classification, the proposed scheme perform a top-down video scene classification which avoids shot clustering. The association for the events of each excitement clip is computed using a priori mining algorithm. A novel sequential association distance to classify the association of the excitement clip into semantic concepts was also proposed in this study. Several semantic concepts for soccer video have been considered ,such as goal scored by team-A, goal scored by team-B, goal saved by team-A, goal saved by team-B.

As mentioned above, there are three problems owing to be further research. (1) The semantic gap problem by directly making use of low-level features for event detection. (2) The rare event training existing in the methods based on statistical models. (3) The application problem of frequent mode association mining method in soccer video. In order to solve these problems, a three-layer framework of semantic event mining for soccer video was proposed in this study, namely shot semantic layer, scene semantic layer and semantic event layer from bottom to top. Furthermore, unlike scene segmentation method used in news video by shot clustering, a new definition about scene in soccer video was also proposed in which the global shot is defined as the boundary of a scene.

THE THREE-LAYER FRAMEWORK OF SEMANTIC EVENT MINING FOR SOCCER VIDEO

The proposed framework is shown in Fig. 1. The shot of soccer video is used as the basic processing unit. A scene semantic layer is added as middle-level between low-level and high-level for bridging the semantic gap. In order to solve the problems of rare event training in the methods based on statistical models and the application

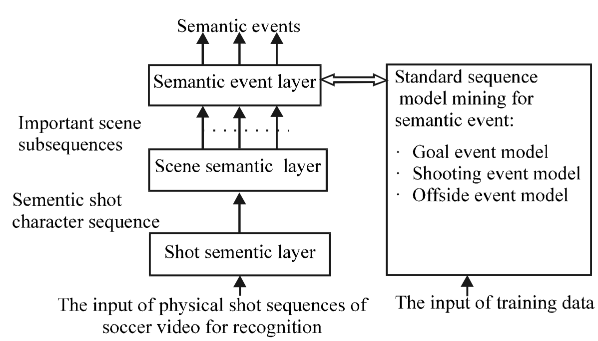

Fig. 1: The semantic event mining framework based on three layer structure for soccer video

problem of frequent mode association mining method in soccer video, at the training stage, the standard sequence model of a known semantic event is mined from the positive instances set by GSP (generalized sequential patterns) algorithm and then event is recognized based on the sequence distance between the standard sequence model of a known semantic event and the frequent mode of a important scene subsequence at semantic event layer.

The original soccer video data is an unstructured successive frame sequence and can be divided into several physical shots by shot segmentation. Shot segmentation method has been at a mature stage and is not in the focus of this study. These physical shots do not contain semantic information, the purpose of shot semantic layer is to recognize the types of them. There are some well-known shot types in soccer video, global shot, medium shot, close-up shot, slow motion playback shot, etc., by which the director and editor show the audience the content of the game. So far there are not uniform standard and method for selecting the shot types which mainly depends on a large number of observations and experimentation. Taking goal event for example, the shot sequence is as follows, global shot→close-up shot of goal player→player's gathering shot→spectator shot→replay shot, etc. We will define and recognize the corresponding shot types as the need of goal and shooting event mining. As the output of the shot semantic layer, the character sequence in which each character stand s for a kind of shot is fed to scene semantic layer for scene segmentation. A scene composes of a successive shots and represents a semantic event occurred in the video. For example, a successive global shots demonstrate the norm play during which no important event is happened while once a goal event occurred, video director will use the shot sequence as mentioned above to show the details of the goal event. A new definition of scene and the corresponding scene segmentation approach for soccer video will be proposed at scene semantic layer. As a result, the character sequence is segmented into scene subsequences which include important scene subsequences and unimportant ones. A important scene subsequences implicates a highlight event and will be fed to semantic event layer for semantic event recognition.

The semantic event layer is composed of two components, i.e., training and recognition. At training stage, the standard sequence model of a known semantic event is mined from the positive instances set by GSP (generalized sequential patterns) algorithm. At recognition stage, the frequent mode of a important scene subsequence is mined by APRIORI algorithm and then event is recognized based on the sequence distance between the standard sequence model of a known semantic event and the frequent mode of a important scene subsequence.

THE REALIZATION THEORY AND METHOD OF EACH LAYER

Shot semantic layer: Two factors should be considered in selecting the shot types for the need of event detection. One is to choose meaningful shot types as much as possible, another is to take the identifiability of shot types into account. At present shot classification methods mostly make use of low-level features, such as color, texture, shape, motion etc. and there are also some practical methods for the detection of players and field lines. In this study, we aim at two kinds of events detection, i.e., goal and shooting. The medium shot therefore, is defined as the player's gathering shot and the close-up shot is further divided into three types, a player shot, a referee shot and spectator shot.

There are totally 7 shot types defined as follows:

- Global shot, represented by character 'A'
- Player's gathering shot, represented by character 'G'
- A player shot, represented by character 'P'
- A referee shot, represented by character 'C'
- Spectator shot, represented by character 'S'
- Replay shot, represented by character 'R'
- And other shot types, represented by character 'O'

Nowadays, there are relative mature methods for these shots recognition. After shot classification, the original physical shot sequence is translated into a character sequence data, e.g., AAAOAAAPA AAAGA…AGRPGA…APRPRA….

Scene semantic layer: There is obvious scene structure like story in news videos or movies, the shots belonging to a scene have the characteristics of visual similarity and time adjacency. Scene segmentation can be realized by shot clustering (Chasanis *et al*., 2007). However, there is not the case in soccer videos, so new definition and segmentation method about scene is needed. Soccer video can be considered as a kind of event-based video. Once a highlight event is occurring, a special shots sequence will be used to show the details. So a semantic event in soccer videos corresponds with successive shots, it is feasible to define and segment scenes taking semantic event as cue. We have observed that successive global shots demonstrate the norm play, during which other single short-time shots occasionally are mixed between them. Once an important semantic event comes up, there must be several other shot types between two global shots used to show the details of the highlight. Based on the observations, we define the global shot as the boundary of a scene and segmented the character sequence into scene subsequences. Also, we have observed that the time of norm play is about above 50% during a soccer match video, so we further divide scenes into two classes. The first class is called unimportant scene which presents the norm play scene that will not be fed to semantic event layer. The second class is called as important scene which starts with a global shot and ends with the next global shot. A important scene subsequence contains a highlight and will be fed to semantic event layer for event recognition.

Semantic event layer: Semantic event layer composes of two components, i.e., training and recognition. The standard sequence mode of a known semantic event is obtained from training data. Sequence distance criterion is used for event recognition:

- The training method of semantic event. A certain number of training samples for a semantic event, of which each sample is a manually labeled scene subsequence and its semantic event is known, are chosen for mining all frequent subsequences whose support is equal or greater than 2 by means of sequence mode mining with GSP (generalized sequential patterns) algorithm. Then we have selected five frequently occurring frequent subsequences for the particular event and determined standard sequence mode in such a way that the length of common subsequence between the standard sequence mode and any member of five frequently occurring frequent subsequences is at least 2
- The recognition of semantic event. Firstly, APRIORI algorithm is used for mining the frequent mode of a important scene subsequence (Zhu *et al.*, 2005; Agrawal and Srikant, 1994). Secondly, the sequence distance between the frequent mode of the important scene subsequence and the standard sequence mode of a known event is computed. Based on the minimum distance criterion the event is recognized, respectively

Definition 1: Set $L_k = \{X_1, X_2,\ldots,X_k\}$ as the frequent mode of a important scene subsequence for recognition, $C_i = \{Y_1, Y_2,\ldots,Y_n\}$ as the standard sequence mode of the ith known semantic event and $X_k, Y_n \in \{A,G,P,C,S,R,O\}$, the distance between L_k and C_i is defined as follows:

$$D\{L_k, C_i\} = 1-\frac{|LCS\{L_k, C_i\}|}{MIN(k,n)}\times\frac{|NCC\{L_k, C_i\}|}{MIN(k,n)}$$

where, LCS demonstrates the length of the longest common subsequence of L_k and C_i, NCC is the number of common characters.

For example, suppose the standard sequence modes of goal event and shooting event are set as C_1 = {P,G,S} and C_2 = {P,R}, respectively, frequent mode L_2 = {G,S} is mined out from the important scene subsequence {A,P,G,S,G,R,S} for recognition by means of APRIORI algorithm, the distances are as following:

$$D\{L_2, C_1\} = 1-\frac{|LCS\{GS,PGS\}|}{MIN(2,3)}\times\frac{|NCC\{GS,PGS\}|}{MIN(2,3)} = 1-\frac{2}{2}\times\frac{2}{2} = 0$$

$$D\{L_2, C_2\} = 1-\frac{|LCS\{GS,PR\}|}{MIN(2,2)}\times\frac{|NCC\{GS,PR\}|}{MIN(2,2)} = 1-\frac{0}{2}\times\frac{0}{2} = 1$$

According to the minimum distance criterion, subsequence {A,P,G,S,G,R,S} can be determined to be a goal event.

RESULTS

In order to verify the performance of the semantic event mining framework presented in this paper, the following experiment program is executed.

Semantic event training: Hundred scene segmentations are used as training samples for each of goal and shooting event, respectively and the shot types of each scene have been manually labeled. Examples are as Fig. 2.

Making use of the method introduced at the section 3, the standard sequence mode of goal event is {PGS} and is {PR} of shooting event.

Semantic event recognition: Five complete soccer videos are used for the recognition of goal and shooting event. The focus of the experiment at this phase is to verify the performance of the semantic event layer, the details of shot semantic layer and scene semantic layer is not considered, so the preprocessing of the five soccer videos still adopts the manually labeling. Each important scene subsequence should be starting with a global shot and ending with the next global shot (this global shot is

Fig. 2: Examples of goal and shooting scenes, A: Global shot, G: Player's gathering shot, P: A player shot, C: A referee shot, S: Spectator shot, R: Replay shot, O: Other shot types

Table 1: Experiment results of goal semantic event recognition

Video number	Name of match	Total time length (min)	Teams	ICL (min)	N_C	N_M	N_F	$N_C/(N_C+N_M)$(%)	$N_C/(N_C+N_F)$(%)
1	Serie A 06/07 season	95	Inter Milan-AC Milan	44	3	0	0	100	100
2	La Liga 06/07 season	96	FC Barcelona-Real Madrid	48	5	0	0	100	100
3	FIFA World Cup 2006	94	Germany-Argentina	38	4	1	0	80	100
4	FIFA World Cup 2006	96	Brazil-France	52	4	0	0	100	100
5	Euro Cup 2008	95	Holland-Romania	34	2	0	0	100	100
Average								96	100

N_C: No. of correctly identified, N_M: No. of miss judgment, N_F: No. of wrong identified, ICL: The time length of important scene, $NC/(N_C+N_M)$: Recall ratio, $N_C/(N_C+N_F)$: Precision ratio, FIFA: Federation International Football Association

not included in this sequence) and will be sent to semantic event layer for recognition. By the way, these important scene subsequences include all occurred highlights of five complete soccer videos, such as goal, shooting, offside, foul, corner, etc.

Because the experiments in this article adopt manually labeled videos, the distance criterion for recognition is defined as: L_k is Ith semantic event if $D\{L_k, C_i\} = 0$, $i = 1$ represents goal semantic event and $i = 2$ represents shooting semantic event; L_k demonstrates other semantic event if $D\{L_k, C_i\} \neq 0$, will not be recognized further. The experiment results are as Table 1 and 2.

As shown in the Table 1, the average recall and precision ratio is above 96%, because the characteristics of scene subsequences representing

Table 2: Experiment results of shooting semantic event

Video number	Name of match	Total time length (min)	Teams	ICL (min)	N_C	N_M	N_F	$N_C/(N_C+N_M)$(%)	$N_C/(N_C+N_F)$(%)
1	Serie A 06/07 Season	95	Inter Milan-AC Milan	44	17	1	2	094	89
2	La Liga 06/07 Season	96	FC Barcelona-real Madrid	48	22	2	5	092	81
3	FIFA World Cup 2006	94	Germany-Argentina	38	14	0	3	100	82
4	FIFA World Cup 2006	96	Brazil-France	52	19	1	4	095	83
5	EuroCup 2008	95	Holland-Romania	34	16	0	2	100	89
Average								096	85

N_c: No. of correctly identified, N_M: No. of miss judgment, N_F: No. of wrong identified, ICL: The time length of important scene, $NC/(N_C+N_M)$: Recall ratio, $N_C/(N_C+N_F)$: Precision ratio, FIFA: Federation International Football Association

goal semantic event is relatively apparent and stable. The close confinement that distance should be 0 results in some individual miss judgment.

In the Table 2, the average recall and precision ratio of shooting semantic event are 96 and 85%, respectively, the main reason is that some foul scene subsequences are very similar with shooting scene subsequence. For example, the subsequence {A, C, P, O, P, R, R} which indicates a serious foul will be mistaken for shooting because its frequent mode is {P, R}. Furthermore, some shooting scene subsequences are omitted because the distance is not 0. For instance, a shooting scene subsequence {A, P, O, R} is miss-judged because the distance between its' frequent mode {A, P, O, R} and standard mode {P, R} is 0.5.

CONCLUSION

The study has presented a three-layer framework of semantic event mining for soccer video, the experiment results for the detection of goal event and shooting event have demonstrated the feasibility of the framework. In scene semantic layer, a new definition about scene in soccer video is proposed, in which the global shot is defined as the boundary of a scene. According to the definition, we segment the semantic shot character sequence into scene subsequences. We have observed that the normal play scenes would account for more than 50% of all scenes in a soccer match video, during which no highlight events occurred, so we further classify scenes into two kinds of important scenes and unimportant scenes. The unimportant scenes do not be sent to the upper layer for recognition with the result that the algorithm time complexity decreases further. The future work is as following:

- The trade off between the selection and identification of shot types should be further considered for a specific semantic event at shot semantic layer
- Scene segmentation approach should be optimized further for the purpose of increasing its anti-interference ability
- More semantic events in soccer video should be involved. In order to improve the recall and precision ratio of the semantic event layer, minimum distance criterion should be adopted, rather than that the distance has been set to 0 in this paper

REFERENCES

Agrawal, R. and R. Srikant, 1994. Fast algorithm for mining association rules in large databases. Proceedings of the 20th International Conference on Very Large Data Bases, September, 12-15, 1994, San Francisco, CA, USA., pp: 487-499.

Ancona, N., G. Cicirelli, A. Branca and A. Distante, 2001. Goal detection in football by using support vector machines for classification. Proceedings of the International Joint Conference on Neural Networks, Volume 1, July 15-19, 2001, Washington, D.C. pp: 611-616.

Assfalg, J., M. Bertini, C. Colombo, A. Bimbo and W. Nunziati, 2004. Semantic annotation of soccer videos: Automatic highlights identification. Comput. Vision Image Understanding, 92: 285-305.

Babaguchi, N., Y. Kawai, T. Ogura and T. Kitahashi, 2004. Personalized abstraction of broadcasted American football video by highlight selection. IEEE Trans. Multimedia, 6: 575-586.

Barnard, M., J.M. Odobez and S. Bengio, 2003. Multi-modal audio-visual event recognition for football analysis. Proceedings of the IEEE 13th Workshop on Neural Networks for Signal Processing, September 17-19, 2003, Toulouse, France, pp: 469-478.

Chasanis, V., A. Likas and N. Galatsanos, 2007. Scene detection in videos using shot clustering and symbolic sequence segmentation. Proceedings of the IEEE 9th Workshop on Multimedia Signal Processing, October 1-3, 2007, Crete, Greece, pp: 87-190.

Ding, Y. and G. Fan, 2007. Segmental hidden markov models for view-based sport video analysis. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, June 17-22, 2007, Minneapolis, USA, pp: 1-8.

Jin, G., L. Tao and G. Xu, 2004. Hidden Markov model based events detection in soccer video. Proceedings of the International Conference on Image Analysis and Recognition, September 29-October 1, 2004, Porto, Portugal, pp: 605-612.

Kolekar, M.H., K. Palaniappan, S. Sengupta and G. Seetharaman, 2009. Semantic concept mining based on hierarchical event detection for soccer video indexing. J. Multimedia, 4: 298-312.

Lefevre, S., B. Maillard and N. Vincent, 2002. 3 classes segmentation for analysis of football audio sequences. Proceedings of the 14th International Conference on Digital Signal Processing, July 1-3, 2002, Santorini, Greece, pp: 975-978.

Li, B., H. Pan and I. Sezan, 2003. A general framework for sports video summarization with its application to soccer. Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, Volume 3, April 6-10, 2003, Hong Kong, China, pp: 169-172.

Li, Y., A. Dore and J. Orwell, 2005. Evaluating the performance of systems for tracking football players and ball. Proceedings of the IEEE Conference on Advanced Video and Signal Based Surveillance, September 15-16, 2005, Como, Italy, pp: 632-637.

Nillius, P., J. Sullivan and S. Carlsson, 2006. Multi-target tracking-linking identities using Bayesian network inference. Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, June 17-22, 2006, New York, USA., pp: 2187-2194.

Ren, R. and J. Jose, 2005. Football video segmentation based on video production strategy. Proceedings of the 27th European Conference on IR Research, March 21-23, 2005, Santiago de Compostela, Spain, pp: 433-446.

Thirumurugan, P. and S.H. Hussain, 2012. Event detection in videos using data mining techniques. Int. J. Comput. Sci. Inform. Technol., 3: 3473-3475.

Wan, K. and C. Xu, 2004. Robust soccer highlight generation with a novel dominant-speech feature extractor. Proceedings of the IEEE International Conference on Multimedia and Expo, Volume 1, June 30, 2004, Taipei City, Taiwan, pp: 591-594.

Wang, L., M. Lew and G. Xu, 2004. Offense based temporal segmentation for event detection in soccer video. Proceedings of the 6th ACM SIGMM International Workshop on Multimedia Information Retrieval, October 10-16, 2004, New York, pp: 259-266.

Xu, M.J. Orwell and G. Jones, 2004. Tracking football players with multiple cameras. Proceedings of the International Conference on Image Processing, Volume 5, October 24-27, 2004, Singapore, pp: 2909-2912.

Yu, T. and Y. Zhang, 2001. Retrieval of video clips using global motion information. Electron. Lett., 37: 893-895.

Zhu X. and X. Wu, 2003. Sequential association mining for video summarization. Proceedings of IEEE International Conference on Multimedia and Expo, Volume 3, July 6-9, 2003, Baltimore, MD., USA , pp: 333-336.

Zhu, X., W.G. Aref, J. Fan, A.C. Catlin and A.K. Elmagarmid, 2003. Medical video mining for efficient database indexing, management and access. Proceedings of the 19th International Conference on Data Engineering, March 5-8 2003, Bangalore, India, pp: 569-580.

Zhu, X., X. Wu, A.K. Elmagarmid, Z. Feng and L. Wu, 2005. Video data mining: Semantic indexing and event detection from the association perspective. IEEE Trans. Knowledge Data Eng., 17: 665-677.