EXTRACTION OF SEMANTIC DESCRIPTION OF EVENTS USING BAYESIAN NETWORKS

[†]*Ahmet Ekin,* [†]*A. Murat Tekalp,* [‡]*Rajiv Mehrotra*

[†]Department of Electrical and Computer Engineering,University of Rochester, Rochester, NY 14627
[‡]Eastman Kodak Company, Research and Development, Rochester, NY 14650
{ekin,tekalp}@ece.rochester.edu, rajiv.mehrotra@kodak.com

ABSTRACT

In this paper, we use Bayesian belief networks to statistically model the trends for event detection. We automatically detect non-rigid object trajectories for object motion units. Then, we use dominant and secondary trajectories of a single object in several consecutive motion units to understand semantic actions or those of more than one object to recognize semantic interactions between objects. We demonstrate sample Bayesian networks to detect events and extract the event descriptions, such as "catch the ball", "throw the ball" and "walk".

1. INTRODUCTION

Visual indexing and retrieval methods using low-level features are not generally suitable for semantic retrieval. Santini and Jain [1] described querying in visual databases as an ill-posed problem and they proposed a solution, staying in the domain of low-level features, using better similarity measures and user feedback. We suggest a semantic model to effectively retrieve high-level information in image and video databases [2].

In still images, extracting objects, understanding the spatial relationships between objects and recognizing the location among several alternatives will be sufficient for most purposes. A Bayesian network based method for semantic object extraction in images was suggested by [3].

For video, semantic object detection can be performed by object segmentation and tracking methods in the literature. Furthermore, the need to detect events in video arises due to the temporal dimension. Multi-agent event detection based on Bayesian networks was proposed by Intille in his Ph.D. thesis [4]. He used Bayesian networks to recognize football plays using object trajectories. The trajectories were drawn by hand and the recognition was based on the selection of the play using the models of the plays in the database. Multi-level event detection using neural networks and inference rules in wildlife documentaries was introduced by [5].

In this paper, we propose a Bayesian network based method to detect semantic video events and extract their descriptions using object trajectories. The trajectories are selected automatically from MPEG-4 sequences within a temporal object motion unit [6]. We select trajectories based on both the dominant and the secondary, but semantically important, motion of objects. The extracted trajectories are input to evidence detectors that also use object reaction units [6] and the decisions about semantic events are given by Bayesian networks. Furthermore, we will also mention using a semantic model for event detection that will help indexing process.

2. BAYESIAN NETWORKS

Bayesian networks are directed acyclic graphs (DAGs) representing the causal dependencies between the nodes that hold variables [7]. Bayesian belief networks are mainly used as a knowledge representation tool in artificial intelligence and expert systems due to their ability to respond to changing conditions easily. Another important property of Bayesian networks is their strength in causal reasoning that is necessary to model actions, explanations, counterfactuals and preferences [8].

The knowledge of the domain is used to construct the network. The inherent uncertainty in the evidences that can be collected from spatio-temporal data set is represented by the prior probabilities of some variables and conditional probabilities between the variables. When an evidence is observed, the evidence is inserted to the network and the posteriori probabilities are calculated using model parameters, priors and conditional probabilities. Certain independence relationships between variables are assumed to exist to efficiently calculate the posteriors when an input variable is observed [7].

The model parameters are set by either training the model or using expert knowledge. The details of the Bayesian networks can be found in [7, 8, 9].

3. LOW-LEVEL EVIDENCES

The low-level evidences are input to Bayesian networks. We use object motion as our primary low-level evidence. The evidences are found for object motion or object reaction units. The motion characteristics of the participants of each unit are presented to the network in the form of trajectory data.

This work was supported by a National Science Foundation SIUCRC grant and a New York Science and Technology Foundation grant to the Center for Electronic Imaging Systems at the University of Rochester.

3.1. Object Motion and Reaction Units

The low-level object motion and reaction units are found according to [6]. Elementary motion units are the temporal segments of object life-span where object motion is coherent. The breaks in the motion generates new motion units. The reaction units are found similarly, but for two objects. The above motion and reaction units do not have semantic meaning. However, the consecutive motion units or reaction units form a semantically meaningful action units for one object, such as "walk", or interaction units for more than one object, such as "throw the ball". Individual action or interaction units and semantic composition of any number of action and interaction units are regarded as events.

3.2. Trajectories

The temporal segments of motion and reaction units form the boundaries for trajectory information. The trajectory data within the segment is fed to the appropriate network for the semantic meaning.

We differentiate two different trajectory types. The first one is the dominant trajectory of the object. The dominant trajectory can be found by calculating the coordinates of the centroid or the same part of the object in every frame. Moreover, a semantic object, if it is non-rigid, will have motion deviating from the dominant motion and the semantically important deviations form the secondary trajectories. We provide a general automatic trajectory extraction given the rough sketch of the object boundaries for the secondary and dominant trajectories.

The dominant trajectory is found by collecting the centroids of the object region for all frames. The method in Figure 1 is used to extract secondary trajectories and it starts with dense motion estimation of the frames in an elementary motion unit followed by the estimation of region (object) motion parameters in terms of 6-parameter affine model [10]. Then, the region affine parameters are compared to individual pixel velocities for each frame in the temporal segment and each individual pixel counter, number of deviations (N_{dev}) from dominant motion within a motion unit, is increased if its velocity value is deviating from the dominant motion. For each pixel, another counter, N_{reg}, is defined to hold the number of frames the pixel is included in the object region in the given motion unit. The consistently deviating pixels are found by thresholding N_{dev}/N_{reg} value for each pixel. The connected regions are found by a labeling algorithm and the velocity and the size of each region is calculated to find important regions. Finally, the centroids of each region for each frame is calculated to give the secondary trajectory data.

Figure 2 shows the results after important steps of the algorithm. The counter image constructed from the region pixels in the temporal motion unit is given in Figure 2(a). The pixels appearing in object region for all frames are shown as the brightest pixels since the counter for them (N_{reg}) is the highest. Figure 2(b) shows the counter image for N_{dev}. The counter for deviating pixels are increased for each frame. However, a pixel may go out of object region for several frames. Thus, the normalization is done by dividing N_{dev} to N_{reg} where N_{reg} is nonzero. After thresholding and filtering out small regions, the regions having enough size and velocity are kept. For the figure, the algorithm keeps only one region that stands for hand motion and eliminates others.

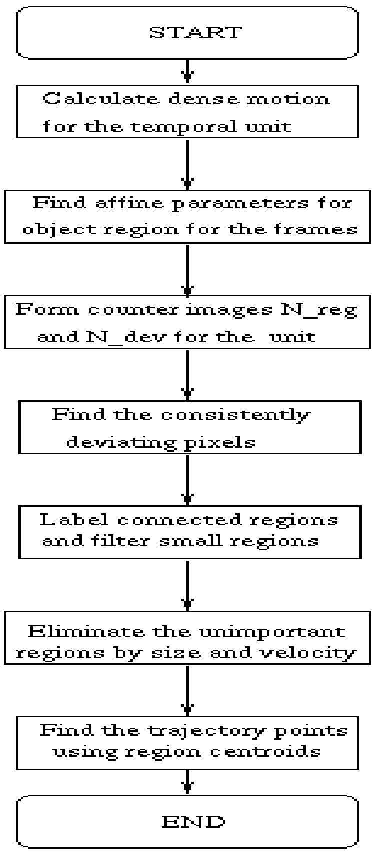
START
Calculate dense motion for the temporal unit
Find affine parameters for object region for the frames
Form counter images N_reg and N_dev for the unit
Find the consistently deviating pixels
Label connected regions and filter small regions
Eliminate the unimportant regions by size and velocity
Find the trajectory points using region centroids
END

Figure 1: The method to find secondary trajectories

4. BAYESIAN NETWORKS FOR EVENT RECOGNITION

Events are composed of object action and interaction units. Action units are defined as semantic object motion units and interaction units are semantically meaningful object reactions. "Walking" and "running" are the events containing only one object. Thus, an action unit of one object may form an event. On the other hand, "throwing the ball" forms an interaction unit between two objects and it is also an event. Finally, "penalty kick" is a composite event consisting of player's "running" action unit and " kicking the ball" interaction unit of the player and the ball.

We constructed three Bayesian networks to detect events and extract their descriptions in Children, Stefan and Hall MPEG-4 sequences. We manually determined the prior and conditional probabilities. However, the use of pre-annotated videos and semantic model can help the automatic adjustment of conditional probabilities.

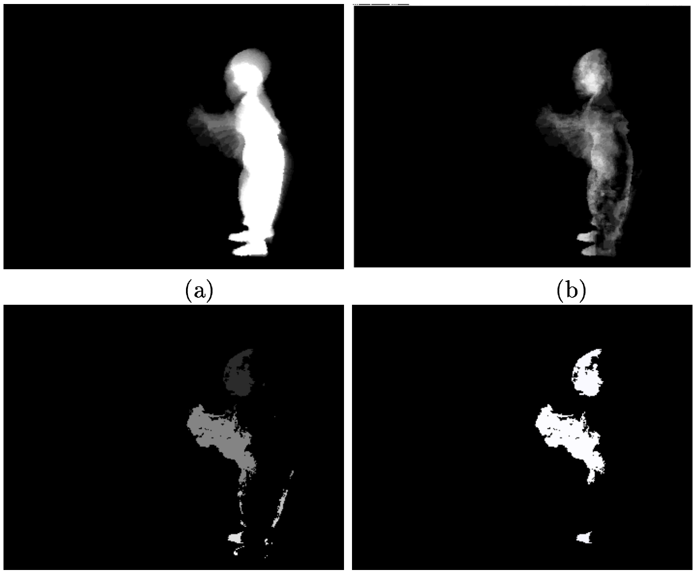
(a) (b) (c) (d)

Figure 2: a)N_{reg} image b) N_{dev} image c) Connected regions d) Small regions eliminated

4.1. Walk, Run, Stationary

The first Bayesian network uses dominant object trajectory to decide whether the object is walking, stationary, running, sitting down or standing up. The network uses only dominant trajectories within object motion units.

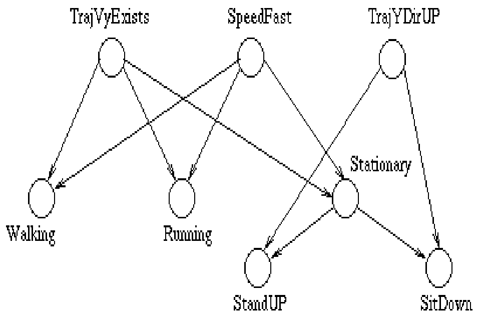

Figure 3: Bayesian network for walk, run and stationary events

The inputs to the network are shown as the first layer in Figure 3: TrajVyExists, SpeedFast and TrajYDirUP. The inputs are assumed to have equal a priori beliefs, such as $P(\text{Vy exists}) = 0.5$ and $P(\text{Vy does not exist}) = 0.5$. The first and third inputs use the projection of the trajectory on the Y axis. Denoting N as the number of points in the trajectory unit and T as the trajectory, we first smooth the trajectory by 3-point averaging to suppress noise effects, then, we find the velocity of the smoothed trajectory along Y axis as in Equation 1.

$$V_y = \sqrt{\frac{\sum_{i=2}^{N}(T.y_i - T.y_{i-1})^2}{N-1}}. \qquad (1)$$

The direction is found by counting the direction change in the smoothed trajectory data and comparing it with the weighted velocity to finalize the decision. The SpeedFast evidence detector calculates the velocity of the smoothed trajectory using both X and Y components. The low-level evidence detectors return three outputs: observed, not observed and not enough evidence. Then, the network uses the observations, prior probabilities and conditional probabilities to find posteriors.

We considered only trajectories to collect evidences, however, the person walking or running towards the camera, such as the second person walking towards the camera in the Hall sequence, cannot be detected using only trajectories. For these circumstances, using the change in the region size may be a solution.

4.2. Hands Up, Down

This Bayesian network uses both types of object trajectories to decide whether hand movement is up or down. The secondary trajectory direction, velocity and the distance between the dominant and secondary trajectories are used as low-level evidences (Figure 4). The low-level evidences SecTyDirUP and SecTSpeedFast use the corresponding algorithms defined in the first network with secondary trajectory data.

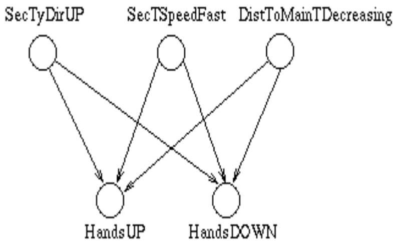

Figure 4: Bayesian network for hands up and hands down

Defining $Dist_x = \sum_{i=1}^{N}(T_{sec}.x_i - T_{main}.x_i)^2$ and $Dist_y = \sum_{i=1}^{N}(T_{sec}.y_i - T_{main}.y_i)^2$, distance to dominant(main) trajectory is calculated as in Equation 2.

$$Distance(T_{sec}, T_{main}) = \sqrt{\frac{Dist_x + Dist_y}{N}} \qquad (2)$$

The network was used to test the secondary trajectory concept and the core parts of the network was transferred to the next Bayesian network shown in a rectangle in Figure 5.

4.3. Catch, Throw or Miss the Ball

The third Bayesian network uses object reaction units, dominant trajectories and secondary trajectories to infer the semantic interactions. Figure 5 shows the Bayesian network designed to detect the catch, throw and miss the ball events. The first layer determines whether object catches the ball or misses the ball depending on the evidences. If there is not enough evidence, the network does not favor any one of the event over the others. "Throwing the ball" event is conditioned on the "catch the ball" event. Temporal sequence of the events are kept in the object reaction unit.

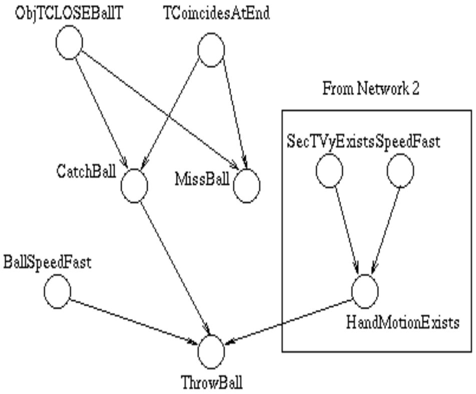

Figure 5: Bayesian network designed to detect catch, throw and miss the ball events

The evidence detectors ObjTCLOSEBallT and TCoincidesAtEnd require some explanation. The first one measures the spatial proximity of object and ball trajectories. The second detector uses the second half of the input segment to decide if the ball trajectory and object trajectory make a similar motion or not.

4.4. Results

The recognition results of the algorithm are given in Table 1. We present the results giving the number of cases network reasoned a correct semantic name, incorrect semantic name or could not reason about the semantic name of an event. The decision is based on the difference in the posterior probabilities. If all of the resulting posterior probabilities are close to each other, the network does not favor any of the events. In our experiment, the first network was able to detect the correct event in most of the cases. The second network was input secondary and dominant trajectories and was able to judge correctly in 14 of 19 experiments. Finally, the third network mislabelled only 1 event out of 9 events.

One of the reasons for missing labels or incorrect labels is the lack of enough data in a motion unit or a reaction unit. The inherent noisy trajectories within a short temporal segment cause false low-level evidences to be inserted to the network. Another reason is the inexact coordinates of secondary trajectories since the automatic extraction of those trajectories is a noisy process. Finally, using only trajectories in some circumstances is not enough to make a decision about an event.

5. CONCLUSION AND FUTURE WORK

Video events were described in terms of action units or interaction units or both. The low-level segments corresponding to object action and interaction units were used to find the object motion trajectories. The trajectories were expressed in terms of dominant and secondary semantically important object motion. The extraction of secondary trajectories were automatically performed. The extracted trajectories were used as low-level evidences to extract event descriptions with Bayesian networks.

Table 1: Results for the Networks

Network Name	Correct Reasoning	False Reasoning	Unable to decide
Walk-Run-Stationary	23/28	4/28	1/28
HandsUPDown	14/19	3/19	2/19
Catch-Miss-Throw	6/9	1/9	2/9

We aim to extend the method using color, texture and shape as low-level evidences in addition to motion trajectories. Furthermore, the possibility of using semantic model to replace the manual adjustment of network parameters will be investigated.

REFERENCES

[1] S. Santini and R. Jain, "Integrated Browsing and Querying for Image Databases", *IEEE Multimedia*, pp. 26-39, July-September 2000.

[2] A. Ekin, A.M. Tekalp and R. Mehrotra, "Object-based motion description: from low-level features to semantics", to appear in *SPIE Storage and Retrieval for Media Databases 2001,* San Jose, CA.

[3] J. Luo, A.E. Savakis, S.P. Etz and A. Singhal, "On the Application of Bayes Networks to Semantic Understanding of Consumer Photographs", *ICIP 2000,* Vancouver, Canada.

[4] S. Intille, "Visual Recognition of Multi-Agent Actions", MIT Ph.D. Thesis, 1999.

[5] N. Haering, R. Qian and I. Sezan, "A Semantic Event Detection Approach and Its Application to Detecting Hunts in Wildlife Video", *IEEE Trans. on Circuits and Systems for Video Tech.,* 10:6, Sep. 2000, pp. 857-868.

[6] A. Ekin, R. Mehrotra and A.M. Tekalp, "Parametric Description of Object Motion Using EMUs", *ICIP 2000,* Vancouver, Canada.

[7] E. Charniak, "Bayesian Networks without Tears", AI Magazine, Winter 1991, pp.50-63.

[8] J. Pearl, *Probabilistic Reasoning in Intelligent Systems,* Morgan Kaufmann, San Francisco, 1988.

[9] F.V. Jensen, *An Introduction to Bayesian Networks,* Springer, New York, 1996.

[10] A.M. Tekalp, *Digital Video Processing*, Prentice Hall, New York , 1995.