BEHAVIOUR PREDICTION USING NEURO-FUZZY HIERARCHICAL ARCHITECTURE

M.SOORYA, S.ABIRAMI
Department of Information Science and Technology
College Of Engineering, Guindy.

ABSTRACT

Behaviour prediction is an interesting research topic which finds applications in many areas like surveillance and security systems. In this paper, a hierarchical architecture is proposed which starts with the lowermost tracking layer followed by neural network and fuzzy layer. The upper layers use the details obtained in the lower layers. The tracking layer outputs the speed and appearance of the object. The position of the objects are found using the ground truth provided with the dataset. The speed values are calculated using the obtained position coordinates. The appearance of the objects are found using person detector which incorporates Histogram Of Gradients (HOG) and Support Vector Machine (SVM). The neural network layer does the pattern recognition task which classifies the micro behaviours like stopping, walking and running based on the speed values obtained from the tracking layer. The fuzzy layer predicts the behaviour by creating the fuzzy inference systems for meeting and fighting using the rule base. Finally the fuzzy outputs are de-fuzzified using the centroid approach. The experiment is evaluated in MATLAB tool using the CAVIAR dataset. Meeting and fighting behaviours are identified with the help of this system.

Keywords— Behaviour Prediction, Neural Network, Fuzzy Inference System

INTRODUCTION

Human behaviour predictions are quite interesting since it has the ability to view the fine grained representation of video (pixels) as semantic behaviour. It gives the system a different view of the video. It addresses many innovative applications in future. Currently it is used in content-based information retrieval, smart video surveillance, ambient intelligence etc... In general, a behaviour prediction system works by extracting the low level features of the video frames and then use machine learning techniques to refine them in a semantic manner. The low level features can be found by finding the trajectories by using optical flow approaches or by tracking otherwise by using the detection algorithms like Mixture Of Gaussians (MOG), Self-Organizing Background Subtraction (SOBS), Enhanced Background Subtraction (EBS) etc. Machine learning approaches which can be used include the supervised and unsupervised learning i.e. either classification approaches like Neural Network, Support Vector Machine (SVM) etc. or clustering approaches. In order to perfectly design a behaviour prediction system it should include pixel level analysis of the video data as well as the context the human activity occurs and also the relationship that exists among the objects at the particular point of time in the scene.

RELATED WORKS

Rosario Di Lascio et al.,[1] explained and compared the performance of algorithms which are used for foreground detection problem. This work, suggests different methodologies for foreground detection like derivative algorithms, background subtraction algorithms and optical flow algorithms. The background subtraction algorithm further classified as probabilistic model, reference model and neural models. The derivative algorithms are classified as single difference algorithm which compares the previous and current frame pixels and the double difference algorithm which consider the variations among more number of adjacent frames. This work consists of the algorithms like the Mixture Of Gaussians (MOG), the Enhanced Background Subtraction (EBS), the Self-Organizing Background Subtraction (SOBS) and the Statistical Background Algorithm (SBA) for the purpose of comparative evaluation. They found that MOG and EBS are effective in most of the cases whereas SOBS perform well in indoor settings.

Giovanni Acampora et al.,[2] explained the hierarchical neuro-fuzzy architecture for human behaviour analysis which incorporates these two approaches to provide tolerance to uncertainties that characterize human behaviour in groups. In this work, a hierarchical approach based on a tracking algorithm, time-delay neural networks and fuzzy inference systems. They introduced a behavioural taxonomy in which the lowest layer is the elementary action followed by the micro behaviour, macro behaviour and group behaviour. The elementary

actions are identified with the help of tracking based layer that consists of the tracking module, post-tracking computation module and context-aware features extraction module. The tracking module gives the position and distance of humans or human and object and also updates the appearance. The post-tracking computation module computes speed and the variation of direction by taking the position input from previous module. The context-aware extraction uses the XML description of environment in which the architecture works. Time Delay Neural Network (TDNN) is used to identify the micro behaviour in which two different classes speed based micro behaviour and loitering based micro behaviour are used. Finally, the fuzzy-based layer which models uncertainty, identifies the macro and group behaviours. It divides the fuzzy architecture of each behaviour as contextual, meeting, left object and danger fuzzy inference systems. When the final value is calculated it is mapped into the fuzzy area and the corresponding behaviour can be found.

K.N.Tran et al.,[3] explained about activity analysis in crowded environments using social cues for group discovery and human interaction modelling which is robust to missed detections, disconnected trajectories and noisy head poses. In this work, a framework is proposed for group activity analysis in which the people in the scene are represented by undirected graph. The vertices are people and the edges are weighted by how much they interact. It uses a two-step process for analysing the group activities (i) discovering the interacting groups based on spatial and orientational relationships between individuals (ii) analysing local interactions in each group to recognize their group activity. Two interaction models are used to find the social interaction cues. The social force model depicts the social interaction area as an ellipse in which small displacement from the centre of the ellipse as the persons head. The social interaction model is explained using the linear, step, power or polygonal function. The Visual focus of attention (VFOA) model depicts the important cues of non-verbal communication to understand the activities of people in groups. Given the head pose, 2D angular view of this person is found by an angular slice centred on the head pose direction. It is based on the following assumptions more closer, more stronger the interaction is and the chance of interaction is higher when two persons facing each other and one person is looking at the other. To discover the interacting group, clustering with dominant set concept is used. Local group activity descriptor is a 2D symmetric histogram of size p*p which helps finding mutual poses of people and their movements in the group. The motion information of people is a very important cue to recognize specific activities. Each video sequence is represented as a collection of LGA descriptor, then the video is represented as a histogram of code words using BOW model. The codebook is constructed by clustering the descriptors using the k-means with Euclidean distance. Finally SVM classifier is used to learn and classify the group activities.

Kang Li et al.,[5] discussed about Prediction of human activity by discovering temporal sequence patterns. In this work, the activities are represented as a sequence of actionlets, which is nothing but a meaningful action unit. The actionlet will have specific semantic meanings and time boundaries. By utilizing this feature, a temporal segmentation method is proposed for identifying the actionlets with the help of motion velocities. Then actionlet categorization is done by using the clustering approach over the actionlet descriptors.
In the representation phase, the activity representation is the main goal. The observation of human activity is temporally segmented into action units using motion velocity as the determining feature. The boundaries between actionlets are detected by finding motion patterns. The observed actionlets are detected and quantized to symbolic labels which map to action classes. In the prediction phase, context information is vital to understand the human activities under particular scene since it typically occurs under particular human and object interaction. So it is understood that, actions as well as the object interactions with actions are needed as a cue. Hence, there is a need for an approach to provide a meaningful prediction with these combinations. This is called eventlet sequences. The context-aware model is built by utilizing sequential pattern mining (SPM) is utilized Apriori-all to incorporate the context cues and the activity which can represent an enriched symbolic sequence.

Rosario Di Lascio et al.,[8] explained the algorithm for tracking people using contextual reasoning. They proposed the method which uses finite state automata for representing objects and maintaining the object information using the states. The states used are new, deleted, frozen, existing, classified, to be classified, in group. When a person enters the scene it is taken as new; When a person exits the frame it is taken as deleted; When a person is inactive it comes under frozen; When a person exits frame but still in the border of the frame it is taken as exiting; When a person completely entered the frame it comes under classified whereas if a person not yet completely entered the frame it is taken as to be classified; When a group is found it is taken as in group. The state manager maintains these states of the objects. By using the history of the states the person can be tracked.

BEHAVIOUR PREDICTION ARCHITECTURE

The design of this system is to build a prediction model for predicting the behaviour in the video sequences. The framework consists of 3 main layers tracking layer, neural network layer and fuzzy layer. Tracking layer tracks the position of the objects in the scene and then compute its speed using the equation. In

the neural network layer, Time delay neural network (TDNN) is used to classify the micro behaviour based on the speed obtained from the previous module. In the fuzzy layer, macro behaviour is found using the fuzzy engine with the help of the contextual rules. Figure 1 depicts the behaviour prediction architecture.

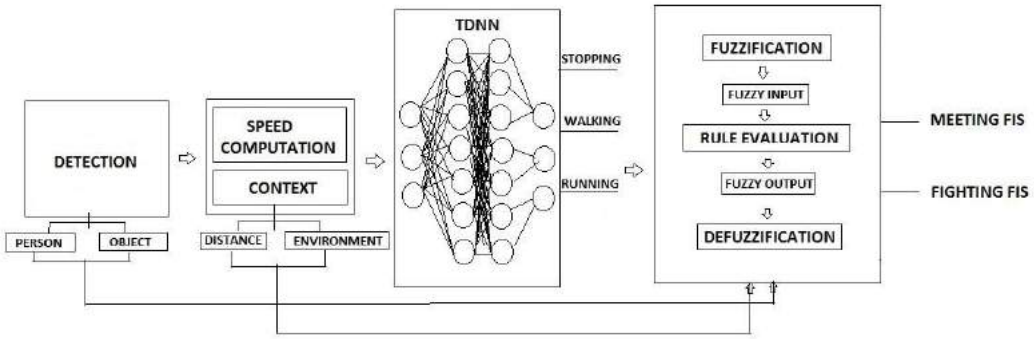

Figure 1: Behaviour prediction architecture.

A. Modules Description

The entire paper is divided into the following modules

- Module 1: Tracking based layer
- Module 2: Post-tracking computation module
- Module 3: Neural network layer
- Module 4: Fuzzy layer

B. Tracking Based Layer

The tracking based layer computes the feature vectors using Histogram of Gradients (HOG) and then classifies the feature vectors using Support Vector Machine (SVM). These algorithms are incorporated in the person detector which finds the upright persons. The position of the persons in the video sequences are obtained from the ground truth data provided with the dataset. The ground truth is given as xml from which the position data is retrieved.

C. Post-Tracking Computation Module

This module takes the position found in the tracking module as input and find the speed using the position obtained from the xml using the following equation

$$\text{Speed}=\sqrt{(x_o^{t} - x_o^{t-\Delta t})^2 + (y_o^{t} - y_o^{t-\Delta t})^2}/\Delta t$$

Where (x_o^{t}, y_o^{t}) is the current position of the object
$(x_o^{t-\Delta t} - y_o^{t-\Delta t})$ is the previous position of the object
Δt is the time period of the scene.

D. Neural Network Layer

The Time Delay Neural Network exploits the temporal characteristic of the data hence suitable for the sequential data like video. Delay is added to the input so as to achieve time-shift invariance. Hidden layers are present between the first and last layer. All the processing are done in these layers with the help of filters. The network should learn optimal filter value at each layer with the help of training sets. The number of neurons in the output will be equivalent to the number of activities. In this paper, the Position, Speed, Variation of direction are given as input which is the sequential data. Pattern recognition task is done using TDNN which classifies the speed based micro behaviours like stopping, walking, running. There are two hidden layers with 15 neurons with the delay of 20 frames and the output consists of 3 neurons.

Steps Involved in Neural Network

- Data collection
- Network creation using the Neural Network Object
- Configuring the network Neural Network Inputs and Outputs
- Initializing the weights and biases
- Training the network
- Validating the network
- Using the network

Back propagation

Learning takes place with the help of back propagation algorithm. Back propagation algorithm works by comparing the neural network's prediction of values with the target value. The goal of backpropagation is to optimize the weights so that the neural network can learn how to correctly map arbitrary inputs to outputs.

E. Fuzzy Layer

The input from the lower layers like the distance, appearance, micro behaviour need to be fuzzified by mapping them to the fuzzy sets by allotting the regions. Now the contextual rules for the pertaining FIS should be defined. The FIS used in this paper are meeting FIS and Fighting FIS. Figure 2 depicts the process involved in fuzzy layer.

Steps Involved in Fuzzy Layer

- Fuzzify the crisp input using the membership function
- Combine the fuzzified inputs with the fuzzy rules
- Find the consequence of this combination from output membership function
- Combine the consequences using the output distribution
- Defuzzify the output using centroid approach

Fuzzification

Fuzzification converts the crisp input to fuzzy input. For example, to classify room temperature as hot and cold. This is achieved with the different types of fuzzifiers (membership functions). The trapezoidal membership function is used in this paper to fuzzify the input. Trapezoidal membership function needs a small amount of data and works well. The condition of a partition of unity is easily satisfied. Figure 3 shows the fuzzification diagrammatically.

Fuzzification

Fuzzification converts the crisp input to fuzzy input. For example, to classify room temperature as hot and cold. This is achieved with the different types of fuzzifiers (membership functions). The trapezoidal membership function is used in this paper to fuzzify the input. Trapezoidal membership function needs a small amount of data and works well. The condition of a partition of unity is easily satisfied. Figure 3 shows the fuzzification diagrammatically.

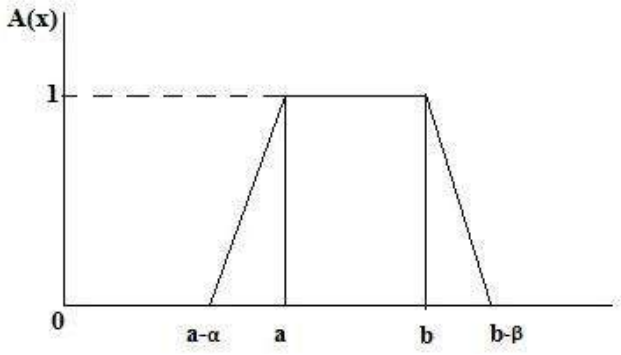

Figure 3: Fuzzification of crisp input.

Defuzzification

Defuzzification is the degree of converting the membership degrees into quantifiable results. Centroid defuzzification approach gives the centre of area under the curve using the equation

$$z^* = \int \mu_B(z).z\,dz / \int \mu_B(z)\,dz$$

Conversion into Crisp Output

The values obtained after the defuzzification is taken and the response is shown as the mapping as to which region the values fall. The semantic label generator generates the corresponding response for meeting and fighting behaviour. Figure 4 shows the sample mapping of the output values where B is the behaviour found using mapping of the values in the output plot. The mapping of the value1 in the meeting area implies meeting behaviour. The mapping of the value2 in the fighting area implies fighting behaviour

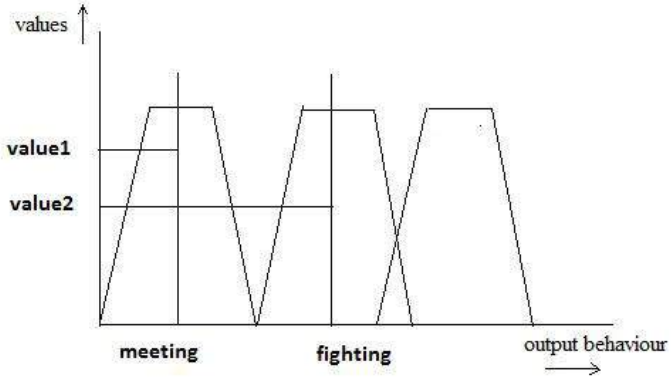

Figure 4: Conversion into Crisp Output.

Contextual Rules

- If distance is small and appearance1 is person and appearance2 is person and stopping1 is yes and stopping2 is yes then Behaviour is Meeting.
- If distance is small and appearance1 is person and appearance2 is person and stopping1 is yes and stopping2 is yes and runningaway is yes then Behaviour is Fighting.

EXPERIMENTAL RESULTS

By using Matlab, the experiments are realized on CAVIAR dataset which consists of a number of video clips for different scenarios like walking, fighting, leaving a package in public. CAVIAR dataset can be found at URL: http://homepages.inf.ed.ac.uk/rbf/CAVIAR/. The video used here is meeting and fighting. For each sequence there are 2 time synchronised videos. The ground truth for the sequences are found by hand labelling the images. The results obtained in each module are discussed as follows. Figure 5.1 shows a snapshot of the input XML file.

```
        <movement evaluation="1.0">walking</movement>
        <role evaluation="1.0">walker</role>
        <context evaluation="1.0">immobile</context>
        <situation evaluation="1.0">moving</situation>
      </hypothesis>
    </hypothesislist>
  </object>
 </objectlist>
 <grouplist/>
</frame>
▼<frame number="25">
 ▼<objectlist>
  ▼<object id="0">
     <orientation>45</orientation>
     <box h="23" w="54" xc="246" yc="275"/>
     <appearance>visible</appearance>
    ▼<hypothesislist>
     ▼<hypothesis evaluation="1.0" id="1" prev="1.0">
        <movement evaluation="1.0">walking</movement>
        <role evaluation="1.0">walker</role>
        <context evaluation="1.0">immobile</context>
        <situation evaluation="1.0">moving</situation>
      </hypothesis>
```

Figure 5: Input XML file.

A. Tracking Based Layer

The tracking based layer computes the speed of the person by accumulating the position of that person in each frame and then computing the speed with the help of the equation. Since in this paper, the focus is on the behavioural analysis, the moving object trajectories mentioned is not used. Instead the ground truth provided with the dataset is used for the purpose of accurate results.

B. TDNN Based Layer

Time delay neural networks (TDNNs) are similar to feed forward networks, except that the input weight has a tap delay line associated with it. This allows the network to have a finite dynamic response to time series input data. This network is also similar to the distributed delay neural network, which has delays on the layer weights in addition to the input weight. Figure 6 shows the time delay neural network. Figure 7 shows the output response. The steps involved in the neural network are as follows.

- Provide the inputs position co-ordinates: x and y, speed
- Multiply weights to the inputs
- Set the delay to 20 frames
- Create two hidden layers with 15 neurons, learning uses back propagation method
- Output will be 3 neurons implying stopping, running and walking

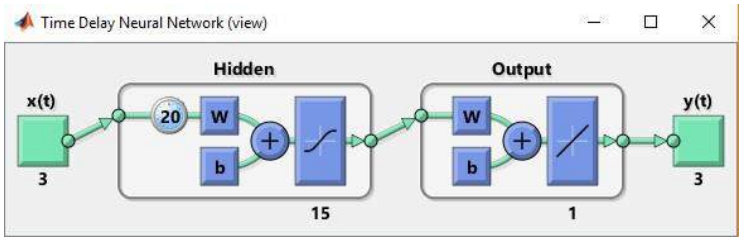

Figure 6: Time Delay Neural Network.

Figure 7: Detection of Micro Behaviour.

C. Fuzzy Based Layer

The aim is to find the behaviour by using the inputs of the previous layers like the distance, appearance and the micro behaviour. The inputs and the outputs are added to the fuzzy inference system followed by the addition of member functions for each input. The member function used here is trapezoidal member function. The rules are added to the system and it can be depicted either as verbose or indexed. The input status words are distance, appearance1, appearance2, stopping1, stopping2, walkingbeforeinactive and runningaway. The output status words are Meeting, Fighting and Browsing. The system is named as fuzzybehaviour FIS. Figure 8 shows the fuzzy inference system created for meeting and fighting behaviour.

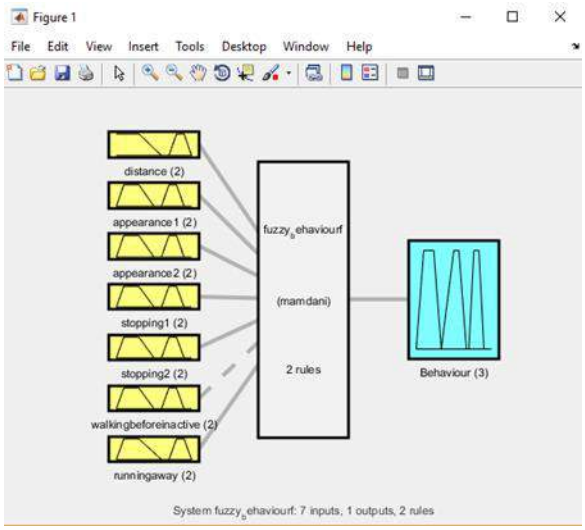

Figure 8: Fuzzy Inference System.

D. Meeting Behaviour

To evaluate the meeting behaviour the first rule mentioned in the rule base must satisfy. It is based on the fact that two objects should be persons and they should walk for a while before stopping for meeting to occur. The distance between them should be small. If it is satisfied then the value will be mapped corresponding to the output of the meeting area computed by fuzzy. Figure 9 depicts the mapping of the output in the meeting area. Figure 10 shows the output response.

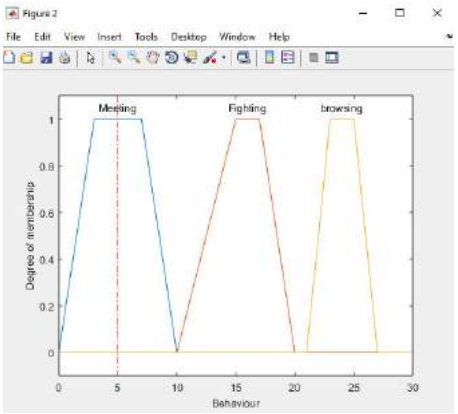

Figure 9: Output of Meeting FIS.

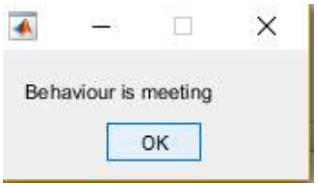

Figure 10: Detection of Behaviour.

E. Fighting Behaviour

To evaluate the fighting behaviour the first rule mentioned in the rule base must satisfy. If it is satisfied then the value will be mapped corresponding to the output of the meeting area computed by fuzzy. Figure 11 depicts the mapping of the output in the fighting area. Figure 12 shows the output response.

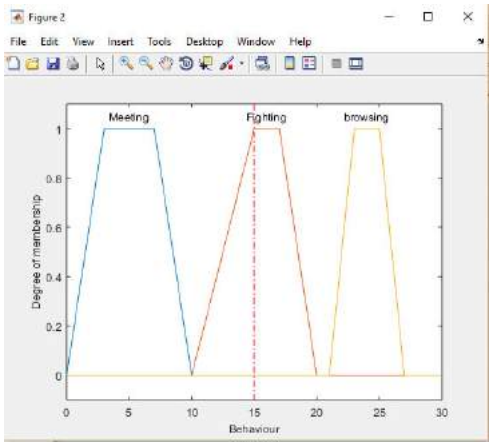

Figure 11: Output of Fighting FIS.

Figure 12: Detection of Behaviour.

CONCLUSION

The aim of this behaviour prediction system is to find the behaviour in the video sequences. The position of the objects are tracked using the tracking based layer. From the position obtained using the tracking the speed of the objects are found. The micro behaviour pertaining to the scene is found using the time-delay neural network. The fuzzy inference system is created for the meeting and the fighting behaviours. Finally, the output is defuzzified to obtain the exact label. The more complex sequences like the behaviour associated with thefts and bomb threats etc. can be found by further improvising this system by generating the xml for the videos while playing and adding the context in future.

References

[1] Conte D, Foggia P, Percannella G, Tufano F and Vento M, "An Experimental Evaluation of Foreground Detection Algorithms in Real Scenes", *EURASIP Journal on Advances in Signal Processing*, vol. 2010, no. 1, pp. 373941, 2010.

[2] Acampora G, Foggia P, Saggese A and Vento M, "A hierarchical neuro-fuzzy architecture for human behaviour analysis", *Elsevier information sciences*, vol. 310, no. c , pp. 130-148, 2015.

[3] Tran K N, Gala A, Kakadiaris I A and Shah S K, "Activity analysis in crowded environments using social cues for group discovery and human interaction modelling", *Elsevier Pattern Recognition Letters*, vol. 44, no. c, pp. 49-57, 2014.

[4] Lao W, Han J and Peter H N, "Automatic video-based human motion analyser for consumer surveillance system", *IEEE Transactions on Consumer Electronics*, vol. 55, no. 2, pp. 59-598, 2009.

[5] Li K and Fu Y, "Prediction of human activity by discovering temporal sequence patterns", *IEEE Transactions on pattern analysis and machine intelligence*, vol. 36, no. 8, pp. 1644-1657, 2014.

[6] Amato A and Lecce W D, "Semantic classification of human behaviours in video surveillance systems", *WSEAS transactions on computers*, vol. 10, no. 2, pp. 343-52, 2011.

[7] Liu R and Zhang X, "Understanding human behaviours with an object functional role perspective for robotics", IEEE transactions on cognitive and developmental systems, vol. 8, no.2, pp. 115-127, 2016.

[8] Di Lascio R, Foggia P, Percannella G, Saggese A and Vento M, "A Real Time Algorithm for People Tracking using Contextual Reasoning", Computer vision and image understanding , vol. 117, no.8, pp. 892-908, 2013.