Human Activity Recognition by Head Movement using Elman Network and Neuro-Markovian Hybrids

Henry C. C. Tan* and Liyanage C. De Silva**

*Department of Electrical & Computer Engineering, National University of Singapore
** Institute of Information Sciences and Technology, Massey University, New Zealand
L.desilva@massey.ac.nz

Abstract

Traditionally, human activity recognition has been achieved mainly by the statistical pattern recognition techniques such as the Nearest Neighbor Rule (NNR), and the state-space methods, e.g. the Hidden Markov Model (HMM). This paper proposes three novel approaches – the use of the Elman Network (EN) and two hybrids of Neural Network (NN) and HMM, i.e. HMM-NN and NN-HMM, to recognize ten simple activities in an office environment. The sex, race and physique invariant feature vectors are extracted from tracking the subjects' head movement over consecutive frames. Based on our database of 200 activity sequences, experimental results show that all the three proposed systems perform better than the two popular conventional methods. The HMM-NN system attained the best performance of 96.5%. The encouraging results not only reveal the performance improvement of combining NN and the traditional HMM, but also demonstrate our proposals' greater potential in realizing recognition of continuous complex activities.

Keywords: Connectionist human activity recognition, human head tracking, Elman partial Recurrent Neural Network, Neural Network and Hidden Markov Model hybrids, digital color image sequences analysis, spatial temporal pattern recognition.

1 Introduction

Human activity recognition (HAR) research has been on the rise because of the rapid technological development of the image-capturing software and hardware, in addition to the omnipresence of reasonably low-cost high-performance personal computers. These new technological advances have made vision-based research much more affordable and efficient than ever before. The main motivation, however, comes from its application in a myriad different challenging but rewarding human-motion-based problems that include automated surveillance systems, human-machine interaction, content-based retrieval, military simulation, clinical gait analysis and sports, etc [1,2]. In all these applications, the ability of the computer vision systems to understand and classify the human activity accurately is very important.

As we have observed, the connectionist techniques, and their hybrids in the form of HMM-NN or NN-HMM, have neither been exploited nor been reported in the literature of HAR. It is thus the objective of this paper to approach the long-standing problem with three solutions based on the artificial neural network, and compare their performance with that of the traditional HAR classifiers – NNR and HMM. In the first proposal, the classifier system based on the Elman model of the partial Recurrent Neural Network (RNN), or simply Elman Network (EN), is advocated. Chosen for its internal representation of time, its ability to remember input from the previous frame and develop an 'understanding' of the context of the input makes it a suitable candidate for the time-varying recognition problem at hand. The second system consists of ten HMMs (each one is trained specifically for a class of activity) and a single-hidden-layer Multi-Layer Perceptron (MLP) NN. The MLP, known for its better classification capability than the HMMs, is used to classify the activity based on the likelihood functions for the ten classes computed by the HMMs. This combination is known as the HMM-NN hybrid. In our final proposal, a NN-HMM hybrid, two MLPs are trained as labelers for ten HMMs, which are time-scale invariant classifiers at sequence level. The MLP, being both trainable and discriminative, is better than the ordinary vector quantizer used in the traditional HMM; hence, this proposed hybrid is also expected to perform better than the traditional HMM classifier.

In all three approaches, we seek to recognize ten distinct classes of activity in an office environment. These activities are walking, squatting, standing up, sitting and getting up, in both lateral and frontal views. For experimental purposes, we built a database of 200 activity sequences, comprising ten activities performed by 20 subjects.

2 System Overview

Our proposed systems are all made up of a few common modules depicted in Figure 1.

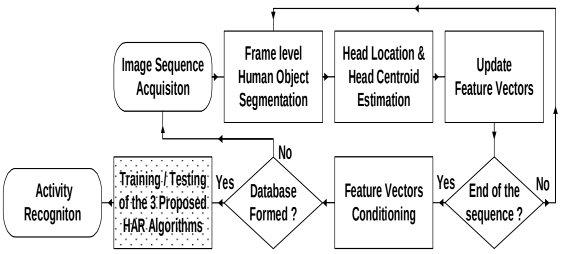


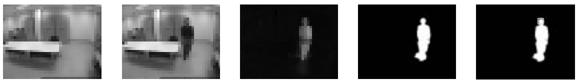
Image Sequence Acquisiton
Frame level Human Object Segmentation
Head Location & Head Centroid Estimation
Update Feature Vectors
No
Activity Recogniton
Training / Testing of the 3 Proposed HAR Algorithms
Yes
Database Formed ?
Feature Vectors Conditioning
Yes
End of the sequence ?
No

Figure 1: System overview of our proposed human activity recognition approaches

Except for the training and testing of the different recognition algorithms, these systems use the same database, have the same preprocessing, feature extraction and feature conditioning modules. Briefly, these modules work together in the following manner. Human activities captured were stored as image sequences. For each frame of every sequence, motion segmentation followed by essential image processing was performed to obtain the human blob. We assumed that our head is always above our torso in the images while performing any of the activities. Then, from the human blob, the head was located automatically and marked with a bounding box. The centroid of the box was used to approximate that of the human head in the frame, and extracted as the feature of interest. By repeating this extraction for all frames in every activity sequence, for every subject, we formed our database. A large portion of the feature vectors was used for the training of each of the classifiers and the remaining 'unseen' samples were subsequently used to test the systems. To improve the efficiency of the limited samples, four-fold cross validation was employed. In the following sections, these processes and their related concepts are explained in greater details.

Many techniques are available for human detection and tracking of moving human. Using either temporal or spatial information of the images, they can be classified roughly into one of the four main approaches. They are statistical motion segmentation method [10], optical flow estimation method [11], motion segmentation method [12], and the background subtraction method [13]. Of these methods, the background subtraction was found to be most computationally efficient and robust enough for our use in the indoor office environment; it was adopted as the means for human motion segmentation in our implementation.

The background, as shown in Figure 2(a), was modeled by computing the mean for each pixel in the colour images over a sequence of 50 frames, which were taken prior to any execution of activity by the actors, e.g. Figure 2(b). Next, the background-subtracted image, Figure 2(c), was subjected to image thresholding, median filtering and some standard morphological operations to segment out the required human blob, as shown in Figure 2(d). The head was then located and marked for subsequent feature extraction (explained in next sub-section), as shown in Figure 2(e). Unlike [8], all these had been accomplished automatically, without the need for human intervention.



(a) background model (b) a frame from a sequence (c) background subtracted (d) morphologically operated (e) head located for feature extraction

Figure 2: Human object segmentation and head location for subsequent feature extraction.

For the experiments, we built a database of 200 human activity sequences – ten different activities performed by 20 subjects, ten from each gender. Subjects are of various height, built and race. Refer to Figure 3 for some snapshots of the recorded sequences for three of the ten activities.

Figure 3: Snapshots of three activity sequences performed by our subjects of different gender, race and physique

3 Proposed HAR Algorithms

In this section, the motivations behind the three proposed approaches are explained and details of their implementation are described.

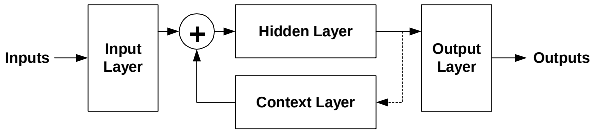


Figure 4: Structure of the Elman Network (EN). (Dotted line represents non-trainable feedback connections.)

3.1 Elman Network (EN)

3.1.1 Motivation of using EN

Our first proposed activity classifier is based on the Elman [18] architecture of the partial RNN, also known as Simple Recurrent Network (SRN) or Elman Network (EN), as shown in Figure 4. The EN is typically employed in situations when we have some data to give to the network for classification, modeling, etc., but the sequence of this input data is important. We want the network to somehow remember the previous inputs and take them into consideration together with the current input data when generating an answer. This memory is achieved by the hidden units feeding its previous outputs back into the context units, which consist of unit delays that store the hidden units' outputs for one time step. All these enable the network to perform learning tasks that extend over time. In fact, it is due to the very nature of the feedback around the hidden units, these hidden neurons continue to recycle information through the network over multiple time steps, and thereby discover the abstract internal representation of time.

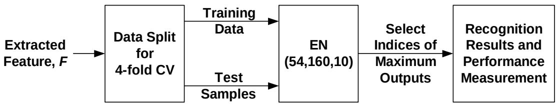


Figure 5: Block diagram of EN-based HAR system.

3.1.2 Applying EN to HAR

The flow chart of our first proposal is shown in Figure 5. As with the traditional classifier systems, the extracted feature matrix was split into four subsets for the training and testing of the EN. The three numbers in the brackets represent the adopted topology of 54-units input layer, 160 hidden-context-units layer and a 10-units output layer – each output unit to represent a class of activity. The number of hidden-context-units was varied from 50 to 500 in the experiment; the value 160 was selected based on the best experimental results obtained. Hyperbolic tangent sigmoid transfer function was used for the hidden layer neurons, and the logistic sigmoid scaling function for the output neurons.

The feedback connection from the output of hidden layer to its input was fixed at 1.0 and the activations of the hidden layer were copied to the context layer on a one-to-one basis. This feedback connection is the dotted line in Figure 4; the solid lines represent trainable connections. The training algorithm fuses the current input with the previous activation of the hidden layer (via the context units) and activates the hidden units with this combined input. The output produced is then compared with a predefined set of desired output. The error generated is used to adjust the strengths of all the trainable connections, so as to move the network outputs closer to the predefined targets; the strength of the feedback connection is left intact. In the training stage, the initial learning rate was set to 0.01, momentum factor to 0.9, all the arbitrary constants to 0.4 and the network is trained by using training data from the three training subsets.

3.2 HMM-NN Hybrid

3.2.1 Motivation of using HMM-NN

In the traditional HMM classifier, each model is trained to maximize the likelihood of producing its training examples but nothing is done to minimize the probability that examples from other classes are produced by the model. This has a negative impact on the recognition capability.

In order to improve the accuracy and retain the endearing time-scale invariant characteristic of the HMM at the same time, our second proposal introduced the incorporation of an MLP at the HMM output.

As the MLP trained as a classifier using the EBP can approximate the Bayes optimal discriminant function and by taking advantage of the discriminative training of the MLP, the weakness in the discrimination ability of maximum likelihood training of the HMM could be overcome. Thus, the recognition performance would be enhanced.

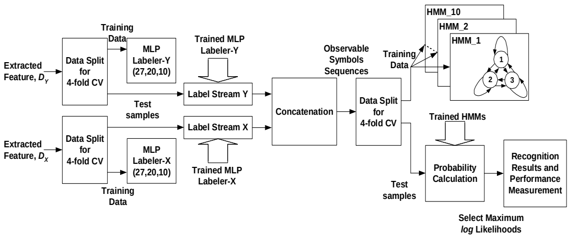


Figure 6: Block diagram of the NN-HMM hybrid system for

3.2.2 Applying HMM-NN hybrid to HAR

For the HMM stage of the hybrid, a three-state ergodic topology identical to traditional HMM classifier systems was used, for ease of comparison. The same training and recognition algorithms were also employed as described in the HMM HAR system. But, instead of assessing the system performance right at the end of HMM stage, its outputs obtained from the Forward algorithm were passed to the input layer of the MLP, and recognition performance was evaluated only at the end of the hybrid system.

For each test sample, the HMM stage output ten *log*-likelihood functions $P(O|\lambda_k)$, where k=1, 2, …, 10. They were then passed to the single-hidden-layered MLP, which has ten input neurons, 50 hidden neurons (this configuration was obtained heuristically, based on the best performance obtained) and ten output neurons, one to represent each class. The number of hidden units was actually varied from ten to 100, by steps of five, in the experiments to obtain the 'optimal' network configuration.

3.3 NN-HMM Hybrid

3.3.1 Motivation of using NN-HMM

In our final proposal, we incorporated two MLPs as labelers for the traditional HMM classifier, resulting in the NN-HMM hybrid. The advantage of such a hybrid system over the traditional HMM classifier is that the MLP, being both trainable and discriminative, outperforms the ordinary vector quantizer and improves the overall recognition capability. The benefit, looking from the MLP point of view, is that HMM will add some dynamic features to the MLP, giving it the capability of handling dynamic HAR problems with the same efficiency and finesse it normally handles static pattern recognition problem.

3.3.2 Applying NN-HMM hybrid to HAR

Two identical MLPs were implemented as labelers for the HMM stage, namely Labeler-Y and Labeler-X (Figure 6). The ten output indices of Labeler-Y were assigned labels '1', '2', …, '10' to represent class '1' to class '10' of our human activity, respectively. Likewise, the output indices of Labeler-X were named '11', '12', …, '20' representing class '1' to class '10', respectively. Each MLP labeler was trained with the modified EBP algorithm to classify vectors for one feature, i.e. either the differences in the x- or the y-coordinates between adjacent frames. The number of hidden units employed in each of the MLP labelers was varied from ten to 100, in steps of five, in the experiments.

To incorporate the MLP output information in the ensuing HMM stage, the straightforward yet effective winner-take-all labeling strategy was applied to the MLP labelers. It took into account the highest scoring output by passing only the label of the top scoring output to the HMM. The HMM then used the resulting label streams as the observation sequences, just as observation symbols from codebooks. Same as the traditional HMM system, the label streams from the MLP were concatenated to facilitate splitting of the data into four subsets for the training and evaluation of the three-state ergodic HMM classifier. As before, one HMM, λ_k (where k=1, 2, …, 10), was trained specifically for each class of activity and training was via the Baum-Welch method of parameter re-estimation that maximized the likelihood function.

4 Results and discussions

On the single assumption that the human head is always above the torso, feature vectors were extracted as described, and a database of 200 activity sequences was built for experimental purposes. For each of the five classifiers, i.e. the two traditional ones and our three proposals, the configuration that gave the 'optimum' system recognition rate was sought for comparison. But by no means are the solutions truly optimal as our searches are not exactly exhaustive. The primary aim of the study is to compare qualitatively our proposals with the traditional means for HAR. So in our opinion, as long as indicative result can be obtained with reasonable amount of resources, and if it allows us to gauge the feasibility of the proposals, it should suffice.

4.1 Recognition using the *k*-NNR

The *k*-NNR classifier was evaluated with *k* taking on values one, three, five, seven and nine nearest neighbors in the experiments. Applying four-fold covaraince values, each test activity sample was assigned to the majority class of its *k* nearest neighbors in the feature space. The estimate of accuracy was obtained from the overall number of correct classifications from all four runs, divided by 200, the total number of samples in the database.

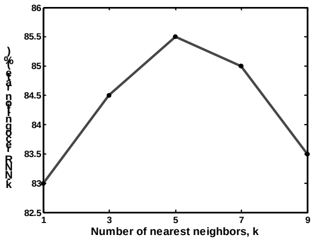


Figure 7: *k*-NNR recognition rate as a function of the number of nearest neighbors used, *k*.

The recognition rate is plotted against the five values of *k* used, as shown in Figure 7. Due to the small data size, the classifier was observed to include more outliers when higher values of *k* were used. The best performance is obtained when *k* is set to five, which has a recognition rate of 85.5%. Thus, the 5-NNR is selected to represent the *k*-NNR method for our HAR classifiers comparison.

4.2 Recognition using the HMM

Since there is no simple theoretically correct way of choosing the number of states, *S*; it was varied from three to ten in the experiment. We fixed the number of symbols *M* at 111 based on the simplified 'quantization' process and used the Forward algorithms to compute the various likelihood functions. The best classification result of 87% is obtained when *S*=3, as shown in Figure 8, the plot of HMM recognition rate versus number of states, *S*. This reveals that in the HMM classifier, the higher number of states does not necessarily imply better performance. On the contrary, a mere three-state model is sufficient to classify our selected human activities, using two one-dimensional sequential features derived from tracking the estimated head centroid (x- and y-coordinates). Hence, the three-state ergodic topology is chosen for the conventional HMM, HMM-NN hybrid and the NN-HMM hybrid classifiers, for easy comparison.

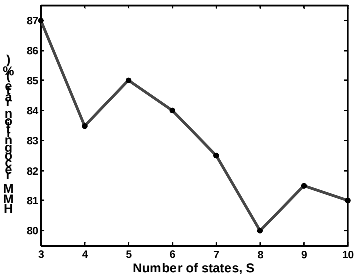


Figure 8: HMM recognition rate as a function of the number of states, *S.*

4.3 Recognition using the EN

In order to find the optimal EN network configuration, different number of hidden units was evaluated. Generally, for an EN to have the best chance at learning a problem, it needs more hidden neurons in its hidden layer than are required for a solution by other method, e.g. the MLP. So, in steps of 50, the number of hidden units was varied from 50 to 500 in the initial attempt to find the 'optimal' network. Figure 8 depicts the plot of recognition rate as a function of the number of hidden units.

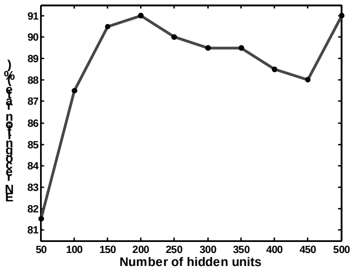


Figure 9: Initial search: EN recognition rate as a function of the number of hidden units.

It was noticed that the EN classifier recognition rate was higher when the network had 150, 200 and 500 hidden units. Despite poorer training performance observed, which sometimes may be good for better generalization of unseen data, the high recognition rate and relatively smaller architecture made the configurations of between 150 to 200 hidden units more attractive than the architecture of 500 hidden units. Hence, it was decided to investigate further and narrow the search for 'optimal' EN configuration to between 150 and 200 hidden units.

4.4 Recognition using the HMM-NN

In this hybrid, in order to obtain the 'optimal' network configuration, the number of hidden units was varied from ten to 100, in steps of five. The recognition rate as a function of the number of MLP hidden units is plotted as shown in Figure 10. The highest performance is achieved when 50 hidden neurons are used in the MLP stage, yielding a recognition rate of 96.5%. The configurations with more than 50 hidden units had probably overfitted the problem and the MLP actually remembered the training examples, resulted in poorer recognition rate. As such, the 50 hidden-units architecture will be used in the HMM-NN classifier for comparison.

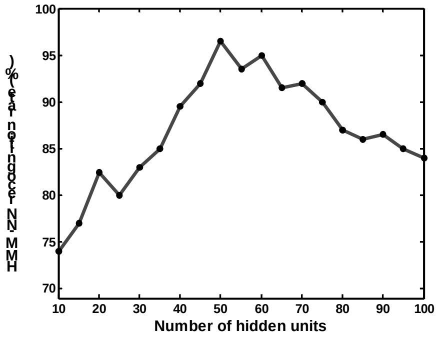


Figure 10: HMM-NN recognition rate as a function of the number of MLP hidden units.

4.5 Recognition using the NN-HMM

Two identical MLPs were implemented as labelers for the HMM stage in this hybrid. Both used the modified EBP algorithm but each was trained to classify vectors for one feature, i.e. either the differences in the x- or the y-coordinates between adjacent frames. The number of hidden units employed in each of the MLP labelers was varied simultaneously from ten to 100, in steps of five. The hybrid system's recognition rate and the labelers' classification rate, both as functions of the number of the hidden units, are plotted in Figure 11. It was noticed that when each of the MLP labelers had 30 hidden units, the best classification results of 96% was obtained at the output of the labelers. However, the best performance of the entire hybrid system does not peak there – it achieved the highest recognition rate of 95% when there are only 20 hidden units in each labeler. The 30-hidden-unit configuration had most likely memorized the training patterns and resulted in inferior overall performance.

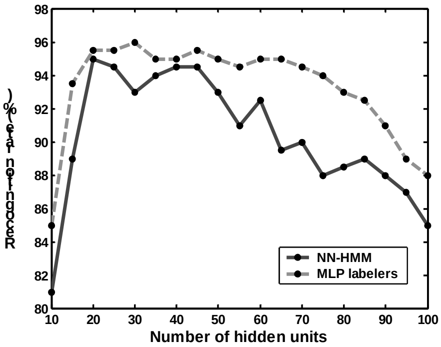


Figure 11: NN-HMM recognition rate and labelers' classification rate as functions of the number of MLP hidden units.

5 Conclusions

This paper presented three novel approaches, based on the EN, the HMM-NN hybrid and the NN-HMM hybrid, for the recognition of ten human activities from a set of color image sequences. The performance of the three proposed systems was evaluated based on a database obtained from motion segmentation, feature extraction and feature vectors

From our investigation, the best recognition rate of 96.5% was achieved by the HMM-NN hybrid. Training of its HMMs was via the Baum-Welch method of parameter re-estimation and its MLP was trained with the modified EBP, whereby approximating the Bayes' classifier. The HMM-NN hybrid's performance is not very much higher than that of the traditional HMM classifier, but it is the most robust candidate for HAR among our three proposals. Nonetheless, the other two proposed systems performed quite well too, attaining recognition rates of 95% (NN-HMM) and 92.5% (EN), as compared to 85.5% and 87% obtained by the traditional *k*-NNR and HMM classifiers, respectively.

6 References

[1] D. M. Gavrila, "The visual analysis of human movement: a survey", *Computer Vision and Image Understanding,* vol. 73(1), pp. 82-98 (1999).

[2] I. Essa, "Computers Seeing People", *AI Magazine*, vol. 20(1), pp. 69-82 (1999).

[3] R. Polana and R. Nelson, "Detecting activities", *Journal of Visual Communication and Image Representation*, vol. 5(2), pp. 172-180 (1994).

[4] X. Sun, C. W. Chen and B. S. Manjunath, "Probabilistic motion parameter models for human activity recognition", *Proceedings of the 16th International Conference on Pattern Recognition*, vol. 1, pp. 443-446 (2002).

[5] J. Ben-Arie, et al., "Human activity recognition using multidimensional indexing", *IEEE Transactions on Pattern Analysis and Machine Intelligence,* vol. 24(8), pp. 1091-1104 (2002).

[6] A. Ali and J. K. Aggarwal, "Segmentation and recognition of continuous human activity", *Proceedings of the IEEE Workshop on Detection and Recognition of Events in Video,* pp. 28-35 (2001).

[7] A. Madabhushi and J. K. Aggarwal, "Using head movement to recognize activity", *Proceedings of the 15th International Conference on Pattern Recognition,* vol. 4, pp. 698-701 (2000).

[8] A. Madabhushi and J. K. Aggarwal, "A Bayesian approach to human activity recognition", *Proceedings of the Second IEEE Workshop on Visual Surveillance*, pp. 25-32 (1999).

[9] T. B. Moeslund and E. Granum, "A survey of computer vision-based human motion capture", *Computer Vision and Image Understanding*, vol. 81(3), pp. 231-268 (2001).

[10] C. Wren, et al., "Pfinder: real-time tracking of the human body", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 19(7), pp. 780-785 (1997).

[11] A. Meygret and M. Thonnat, "Segmentation of optical flow and 3D data for the interpretation of mobile objects", *Proceedings of the Third International Conference on Computer Vision*, pp. 238-245 (1990).

[12] S. Huwer and H. Niemann, "Adaptive change detection for real-time surveillance applications", *Proceedings of the Third IEEE International Workshop on Visual Surveillance*, pp. 37-46 (2000).

[13] M. Seki, H. Fujiwara and K. Sumi, "A robust background subtraction method for changing background", *Proceedings of the Fifth IEEE International Workshop on Applications of Computer Vision,* pp. 207-213 (2000).

[14] J. K. Aggarwal and Q. Cai, "Human motion analysis: a review", *Computer Vision and Image Understanding,* vol. 73(3), pp. 428-440 (1999).

[15] R. Kohavi, "A study of cross-validation and bootstrap for accuracy estimation and model selection", *Proceedings of the 14th International Joint Conference on Artificial Intelligence*, pp. 1137-1143 (1995).

[16] L. R. Rabiner and B. H. Juang, "An intr. to Hidden Markov Models", *IEEE ASSP Magazine,* vol. 3(1), pp. 4-6 (1986).

[17] J. Yamato, J. Ohya and K. Ishii, "Recognizing human action in time-sequential images using Hidden Markov Model", *Proceedings of the Computer Vision and Pattern Recognition* , pp. 379-385 (1992).

[18] J. L. Elman, "Finding structure in time", *Cognitive Science*, vol. 14, pp. 179-211 (1990).