Comparison of Human and Machine Recognition of Everyday Human Actions

Trevor D. Jones[1], Shaun W. Lawson[1], David Benyon[2], and Alistair Armitage[2]

[1] Lincoln Social Computing (LiSC) Research Centre,
Department of Computing and Informatics,
University of Lincoln,
Brayford Pool, Lincoln, UK. LN6 7TS
{tjones,slawson}@lincoln.ac.uk
[2] School of Computing,
Napier University, 10 Colinton Road,
Edinburgh, UK, EH10 5DT
{d.benyon,a.armitage}@napier.ac.uk

Abstract. The research presented here makes a contribution to the understanding of the recognition of biological motion by comparing human recognition of a set of everyday gestures and motions with machine interpretation of the same dataset. Our reasoning is that analysis of any differences and/or correlations between the two could reveal insights into how humans themselves perceive motion and hint at the most important cues that artificial classifiers should be using to perform such a task. We captured biological motion data from human participants engaged in a number of everyday activities, such as walking, running and waving, and then built two artificial classifiers (a Finite State Machine and a multi-layer perceptron artificial neural network, ANN) which were capable of discriminating between activities. We then compared the accuracy of these classifiers with the abilities of a group of human observers to interpret the same activities when they were presented as moving light displays (MLDs). Our results suggest that machine recognition with ANNs is not only comparable to human levels of recognition but can exceed it in some instances.

Keywords: Neural network, finite state machine, moving light display, human biological motion.

1 Introduction

Gunnar Johansson [1] first illustrated how humans are skilled at visually analysing and recognising human motion from very sparse and impoverished datasets – or moving light-displays (MLDs). Since Johansson's pioneering work a great deal of literature has appeared on the subject of so-called biological motion – though an exact understanding of how humans understand MLDs has yet to be reached. In parallel with such work, the computer vision community has produced a wealth of approaches to segmentation of the spatio-temporal information held in video images as well as the classification of the feature sets determined therein (for reviews of such work see, for

V.G. Duffy (Ed.): Digital Human Modeling, HCII 2007, LNCS 4561, pp. 120–129, 2007.

instance, Gavrila [2] and Essa [3]). In general, the range of approaches adopted has varied enormously depending on the application domain and its constraints. Many different pattern recognition methods have been applied to the problem including artificial neural networks (ANNs), statistical classifiers and rule based systems. Recently some studies have appeared which attempt to exploit what is known about human recognition of biological motion to inspire the development of autonomous machine recognition of the same phenomena (e.g. [4]). However, even here there is debate over whether a model based approach to the problem, which uses prior knowledge, such as a model of physical make-up of a human body, to assist classification should be favoured over classification of 'raw' spatio-temporal data.

The research presented here makes a contribution to the understanding of the recognition of biological motion by comparing human recognition of a set of everyday gestures and motions, presented as MLDs, with machine interpretation of the same dataset. Our reasoning is that careful analysis of any differences and/or correlations between the two could reveal insights into how humans themselves perceive motion and hint at the most important cues we use for this. For instance if a machine classifier can accurately recognise an action as well as a human, without any programmed knowledge of a human model or any other background or information (or context), then we could assume that we only need the information within the MLD to interpret the action. The motivation for our work is to try and understand how we might build interactive computer systems that can simulate a natural understanding of biological motion. In particular, and in the longer term, we are interested in seeking an understanding of how computer systems might be used to detect subtle changes in biological motion which indicate changes, for instance, in a human's health or emotional state.

The following section briefly describes some of the artificial classifier techniques used to classify human biological motion and problems arising from popular techniques. This is followed by an account of human biological motion in the form of MLDs. We then present our experimental work, results and conclusions.

2 Classification Techniques

Artificial neural networks are a popular means of classifier, they are well suited for appearance based representations and applications that benefit from fast response times [5]. ANNs are trained from data samples without the requirement for explicit data modelling using a weighting algorithm [6], and do not require large memory storage. The most commonly used neural network is the multiplayer perceptron [7] though other networks such as SOM (Self Organising Maps), or Kohonen map, are favoured for visual recognition systems [8].

Rule based classifiers such as finite state machines are a means of tracking transitions through sequences known as states. Each state indicates a possible path of execution reflecting input changes within the system with a transition in state occurring once a condition has been fulfilled [9]. The finite state machine (FSM) segments events into key stages that represent events within the system. It is useful for systems that have a definable flow and change according to triggered events. The diverse ways and lack of uniformity in human movement, coupled with possible

signal noise, is enough to prevent a state being activated. These findings were an issue for [10] who use an FSM to track hand and head gestures, though the FSM approach does fair well with movements that are less complex and have little diversity [11].

Statistical models such as the HMM (Hidden Markov Model) have some limitations in representing non-gesture patterns [12], but are useful for modelling spatio-temporal variability where the data is temporally well aligned. Similar to the finite state machine, HMMs use state transition but with probabilistic algorithms. Another way of distinguishing behaviour from movement is using a database of possible movements for each action [13]. This type of system is reliant on predefined constraints that would require the creation of extensive profile sets for each participant. Probabilistic networks have been used with some success with recognising small movements such as those performed with the head [14] but have had improved accuracy recognising human activity when used in conjunction with a hidden markov model [15].

The problems with recognising gesture are numerous: pattern spotting [16] – locating meaningful patterns from an input data stream. Segmentation ambiguity [17] - where does the gesture begin and where does the gesture end? Spatio-temporal variability [18] – gestures vary dynamically in shape and duration, be it a large number of people or just one person making the gestures. Vision based systems suffer from their poor ability to continuously track the object of focus, mainly the hands and face. When the focus is lost, perhaps through tracking error or occlusion, the system must be recalibrated requiring the object of focus (the person) to stand in a default pose for the system to once again acquire tracking [19]. A vision based approach also suffers from location limitation – it can only track within its field of vision. It does however, allow for unencumbered freedom for those that are being tracked. A sensor based approach, on the other hand, does not suffer from occlusion, lighting conditions and other visual constraints; it can track a person's movement continuously. However, sensors are invasive and can be cumbersome.

3 Human Biological Motion and Moving Light Displays

Humans can quickly detect other humans within view and, in most cases, can determine biological aspects such as their sex, approximate age, posture, gait, weight, height and build. We recognise social significances that lead us to approximate health, strength, physical ability, and from a persons bearing can make assumptions about intent: aggressive, placid, gentle, untrustworthy, trusting etc. Human motion contains a wealth of information: actions, traits, emotions, and intentions. Our ability to extract complex visual information from the perceived world has been widely documented [20] and theorems for understanding environmental and contextual cues. Gunnar Johansson demonstrated the ability of humans to visually analyse and recognise human motion from very sparse and impoverished datasets. Information of the human form, as represented by an MLD, is considerably reduced (Fig 1). When an MLD is stationary the information presented is near meaningless, but when the MLD is in motion an observer is able to perceive the definition of a human form [1].

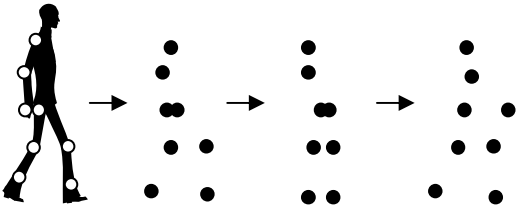


Fig. 1. An MLD representing a person walking

Experiments using MLDs have shown that humans are able to, within some degree, recognise not only the movement but also whether the form is male or female [21] and even to recognise themselves and people familiar to them [22]. Runeson and Frykholm [23] attempted to disguise the gender of the person in the point light display by asking actors to perform as though they were of the opposite sex; observers guessed the correct gender of the actor over 85% of the time. Runeson and Frykholm also showed actors throwing sandbags out to varying distances, the actor was fitted with the lights but the sandbags were not. Observers were good at judging how far the bags would have been thrown.

It is not only humans that can recognise biological motion in MLDs. In their experiment with newly hatched chicks Vallortigara et al [24], showed animated images of 'point light hens' to chicks which would be the first visual stimuli they would encounter after hatching. The chicks followed the motion of the point light hens showing that they perceived the motion. They then performed further tests to answer whether the chicks' response was innate or a learned experience. This time they used artificially induced motion patterns but the chicks were still drawn to the biological motion of not just hens, but also that of cats. Other types of models were used which used rigid structures to define body shape or models which represented a hen-like object, but again the chicks were attracted to the biological motion of the point light hen. As a control, they showed the chicks point light models of both hens and cats, for which the chicks were just as likely to approach the cats as they were the hens. They suggest that from these results chicks have evolved a predisposition to notice objects that move like vertebrates, which may maximise the probability of imprinting on the object most likely to provide food and protection after birth. They refer also to similar findings in four month old human babies and conclude that the results suggest that this preference is hard wired in the vertebrate brain.

4 Experimental Procedure

In this work our intention was to compare machine interpretation of biological motion with human interpretation of the same data. We captured biological motion data from human participants engaged in a number of everyday activities, such as walking, running and waving, and then built two artificial classifiers (a Finite State Machine and a multi-layer perceptron artificial neural network, ANN) which were capable of discriminating between activities. The intention was then to compare the accuracy of

these classifiers with the abilities of a group of human observers to interpret the same activities. A sensor based approach was chosen over a visual recognition system to capture human motion, which took place in a small gymnasium. To capture walking and jogging actions the participants were required to walk and jog on a commercial running machine. Punching was performed by hitting a punchbag and throwing was performed by throwing tennis ball sized balls. Data from typical everyday full-bodied human behaviours were collected using a commercial sensor system (Polhemus Liberty) usually used for motion-capture animation work [25]. We used just five sensors – forty human participants were recruited and sensors placed on: each hand, each foot and the forehead. Participants were asked to perform the following motions: walking, jogging, shaking-hands, waving, punching, kicking, throwing a ball, looking left/right, looking up/down. During each motion, spatial (x,y,z) data was recorded for each participant. Our resultant dataset was therefore very rich in terms of the breadth of examples we recorded – however we deliberately kept the content of each recorded instance of motion very sparse and impoverished – with only 5 points recorded. Two commonly used autonomous classifiers were then trained and configured to classify the data, a Finite State Machine (FSM) and an Artificial Neural Network (ANN). The data is then converted into MLDs for classification by forty human participants.

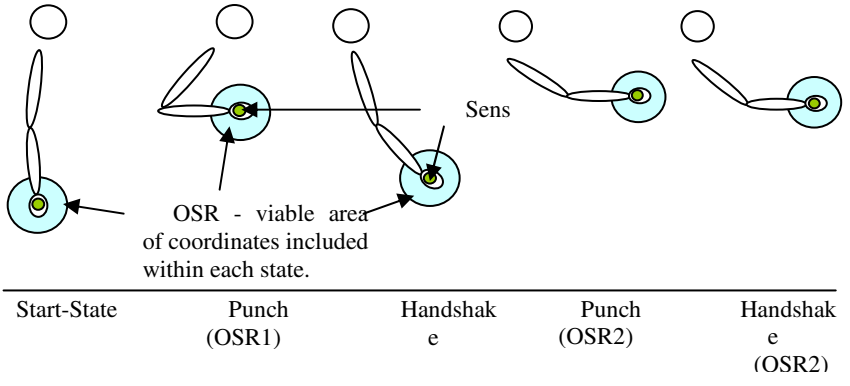


Fig. 2. FSM Optimal state ranges (OSR) for a punch and a handshake

4.1 Finite State Machine

Firstly, an FSM was configured to classify actions; this method hence combined both human and autonomous data perception (since human knowledge of the raw sensor data was incorporated into the design of the FSM). The FSM was designed to calibrate for each participant e.g. measure the length of the participants arm and setting the distance value for the states. For example, if the states were calibrated for a person with long arms the state activation area would be too far to away to be activated by a person with shorter arms. The FSM has four states: a start state, a predictive state, a hit state and a null state. The first three states comprise of an OSR (optimal state range): a spatial three dimensional sphere-shape in space which moves with the participant illustrated in Fig 2.

4.2 Artificial Neural Network

Secondly, a multi-layer-perceptron ANN was trained to classify the raw sensor data - i.e. no model of a human, or any other prior real world knowledge, was integrated into the training phase. The capture data, for all gestures, was re-sampled using linear duplication to make it temporally aligned and then converted for suitability with Trajan, a commercial ANN application (Trajan, date). The data was pre-process using PCS (Principal Component Analyses). PCA is a linear dimensionality reduction technique, which is able to identify orthogonal directions of maximum variance and project the data into a lower-dimensionality space formed of a sub-set of the highest-variance components [26]. The data was sampled using evenly distributed subsets for test, training and verification.

4.3 Human Classification

Additionally, our motion capture data was converted into MLDs by importing it into a 3D animation package and creating short video clips showing five 3D spheres which move in accordance with the captured data on a plain background. We then recruited a further forty human participants who viewed the simulated MLDs, watching each gesture in turn and stating the full-body gesture that they thought was being exhibited. The classification capabilities of the autonomous systems were then compared to those of the human participants. Motion capture data was converted into MLDs by importing the data into a 3D animation package and mapping spheres to the Cartesian coordinates, a representative sample of which is presented in Fig 3. Forty people from varying disciplines and backgrounds participated in the experiment, male and female ranging from ages eighteen to fifty eight.

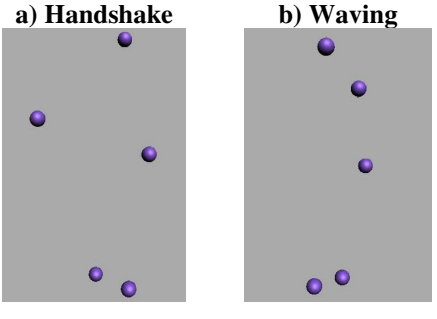


Fig. 3. a) Static image of person b) Snapshot of a person waving

The experiment comprised of three stages: Part 1: a participant viewed a static image of the spheres representing a person (upright and with their arms by their sides). They were asked to give their first impressions of what came to mind when they saw the image. Part 2: The participant was shown an animated MLD for all of the gestures, and for each, was asked to state what they thought they were seeing. Part 3: the participants are informed the spheres represent human motion and are asked to view the MLDs once again. Part 4: All of the MLDs were compiled, in a different

order, into one animation medley. Random movement of the spheres was placed between actions signifying other types of movement or ambiguous action. The participants are informed of which animations represent which action and are then shown an animation medley compiled from were asked to say aloud the gesture they recognised.

5 Results

The FSM was able to recognise gestures with less variability of motion than those which can be performed with greater exertion (Fig 4). Due to the rigid OSR structure the FSM will only recognise movement that keeps within a spatial boundary relative to a fixed point. Gestures which have the potential for greater exertion e.g. punching and throwing were often performed beyond the OSR limits. Recognition for walking and jogging was considerably lower than other results mainly as a result of fluctuations in the sensor output signal registering outside the OSR. This has potential for improvement through redefining the experimental design.

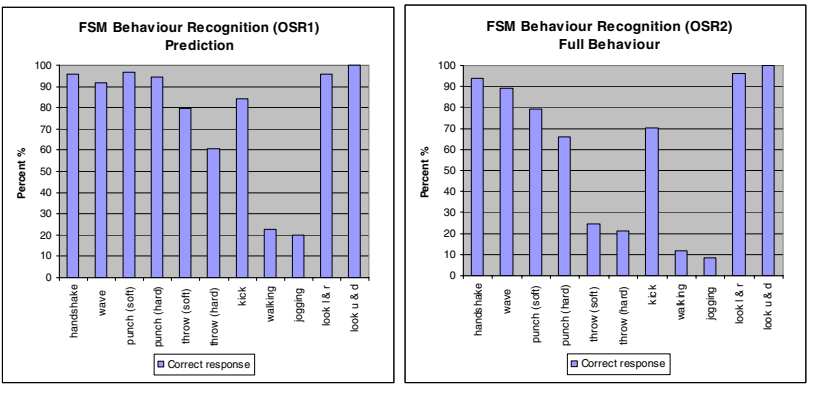


Fig. 4. Results for OSR1 and OSR2 Recognition

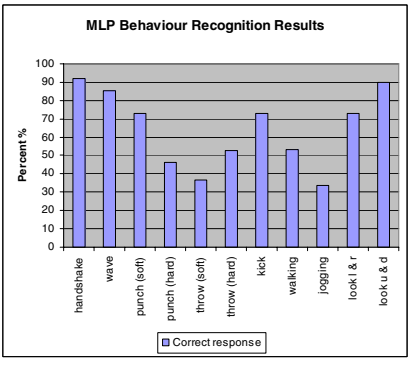
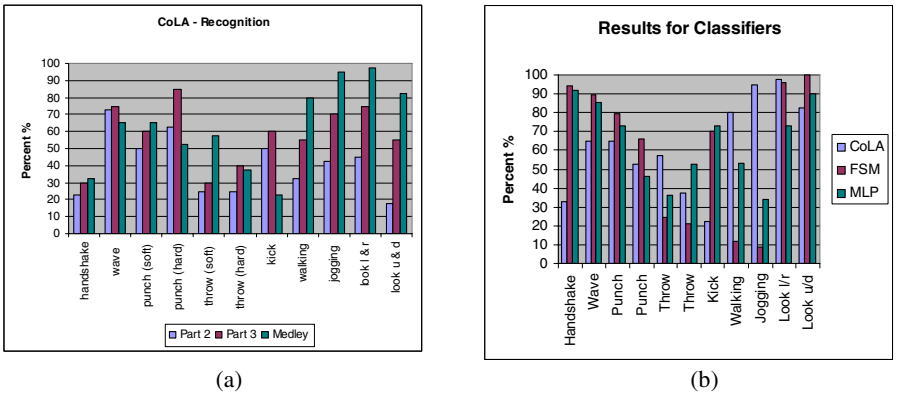


Fig. 5. Results for MLP

The MLP classification performance suffered most of all with actions such as walking, jogging, punching and throwing; for the arm actions there is more variability in these movements than the actions the MLP had less trouble classifying (Fig 5). The walking and jogging may also have posed some difficulty as there was a considerable amount of variance in the data. This variance is due to the participants' feet impacting on the running machine with each stride; the sensor effectively sustains excess motion from vibration. The difference in individual style of performing certain gestures may also be one possible reason for low recognition.

When human participants were shown the static MLD their recognition of the spheres representing a static human was low, 45% in comparison to when the animations were in motion 82%. Johansson makes this observation noting observers rarely recognise the human form in static displays [1]. The results for part three (Fig 6a) would suggest that the increase in some of the recognition results from part two is primarily due to the introduction of partial-context (knowledge the MLD represents human form), and also that of training effect. Recognition is poor for punching and throwing hard, these gestures are quite erratic when performed with pronounced physical amplitude having an adverse affect on the spatial pattern structure. The recognition for handshake was consistently poor; there are a greater number of actions performed in front of the body e.g. shaking hands, turning key in lock, manipulation of objects etc. than those performed at the sides or high up in relation to the torso e.g. waving. The viewing angle of the MLD may have contributed to low recognition of some gestures such as the kick which may need a more prominent perspective. Overall, across all actions, the results of our three classification approaches were: FSM 60.08 %, MLP 64.45 %, human 62.5 % (Fig 6b).



(a) (b)

Fig. 6. (a) Results for parts 2, 3 and animation medley (b) Results for Classifiers

6 Summary of Conclusions

Our results suggest that, at our level of abstraction, machine recognition with ANNs is not only comparable to human levels of recognition but can exceed it in some instances (for instance, in our results, the ANN and FSM showed superior

classification of waving and hand shaking). However, we also found that humans were good at interpreting head and some hand gestures but particularly gestures involving the feet. It is suggested that absence of contextual cues is the main reason for the lower performance in human recognition. Additionally, our work allowed the MLP to utilise the full 3D nature of our data whilst humans were only able to view 2D projections of this data as a MLD. However overall, we believe that the finding that autonomous classifiers which make no use of prior real world information (such as a human model) can potentially perform recognition of biological motion as well as a human is a significant finding which warrants further investigation.

References

1. Johansson, G.: Visual perception of biological motion and a model for its analysis. Perception and Psychophysics 4(20), 201–211 (1973)
2. Gavrila, D.M.: The Visual Analysis of Human Movement: A Survey. Computer Vision and Image Understanding 73, 82–98 (1999)
3. Essa, I.A.: Computers Seeing People. AI Magazine 20, 69–82 (1999)
4. Laxmi, V., Carter, J.N., Damper, R.I.: Biologically-Inspired Human Motion Detection. In: Proc of ESANN - European Symposium on Artificial Neural Networks, Bruges, Belgium, 24-26, 2002, pp. 95–100 (2002)
5. Van Laerhoven, K., Aidoo, K.A., Lowette, S.: Real-time Analysis of Data from Many Sensors with Neural Networks. In: 5th IEEE International Symposium on Wearable Computers (ISWC) 2001, vol. 115 (2001)
6. Heidemann, G., Bekel, H., Bax, I., Saalbach, A.: Hand Gesture Recognition: Self-Organising Maps as a Graphical User Interface for the Partitioning of Large Training Data Sets, icpr. In: 17th International Conference on Pattern Recognition (ICPR'04), 2004, vol. 4, pp. 487–490 (2004)
7. Rumelhart, D.E., McClelland, J.: Parallel Distributed Processing. vol. 1, MIT Press, Cambridge, MA (1986)
8. Corradini, A., Gross, M.: A Hybrid Stochastic-Connectionist Architecture for Gesture Recognition. In: International Conference on Information Intelligence and Systems 1999, p. 336 (1999)
9. Hong, P., Turk, M., Huang, T.S.: Constructing Finite State Machines for Fast Gesture Recognition. In: ICPR '00: Proceedings of the International Conference on Pattern Recognition 2000, IEEE Computer Society (2000)
10. Hong, P., Turk, M., Huang, T.S.: Gesture Modelling and Recognition Using Finite State Machines. In: IEEE, Fourth International Conference on Automatic Face and Gesture Recognition 2000, pp. 28–30 (2000)
11. El Kaliouby, R., Robinson, P.: Real Time Head Gesture Recognition in Affective Interfaces. Human-Computer Interaction Journal '03, 950–953 (2003)
12. Lee, H., Kim, J.H.: An {HMM}-Based Threshold Model Approach for Gesture Recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence 21, 961–973 (1999)
13. Emering, L., Boulic, R., Thalmann, D.: Interacting with Virtual Humans through Body Actions. IEEE, Computer Graphics and Applications 18, 8–11 (1998)
14. Madabhushi, A., Aggarwal, J.K.: A Bayesian Approach to Human Activity Recognition. In: Second IEEE Workshop on Visual Surveillance 1999, p. 25 (1999)

15. Sun, X., Chen, C., Manjunath, B.S.: Probabilistic Motion Parameter Models for Human Activity Recognition. In: Proceedings of IEEE International Conference on Pattern Recognition (ICPR) Québec City, Canada,2002 (2002)
16. Rose, R.C.: Discriminant Word Spotting Techniques for Rejection Non-Vocabulary Utterances in Unconstrained Speech. In: Proc. IEEE Int'l Conf. Acoustics, Speech, and Signal Processing, San Francisco, 1992, vol. II, pp. 105–108 (1992)
17. Takahashi, K., Seki, S., Oka, R.: Spotting Recognition of Human Gestures from Motion Images Technical Report IE92-134, The Inst, of Electronics, Information, and Comm. Engineers, Japan, pp. 9–16 (1992)
18. Baudel, T., Beaudouin-Lafon, M.: CHARADE: Remote Control of Objects Using Free-Hand Gestures. Commun. ACM 36(7), 28–35 (1993)
19. Nickel, K., Stiefelhagen, R.: Pointing gesture recognition based on 3D-tracking of face, hands and head orientation. In: 5th international conference on Multimodal interfaces, 2003, pp. 140–146 (2003)
20. Desolneux, A., Moisan, L., More, J.: Computational gestalts and perception thresholds. Journal of Physiology - Paris 97, 311–324 (2003)
21. Kozlowski, L.T., Cutting, J.E.: Recognising the gender of walkers from point-lights mounted on ankles: Some second thoughts. Perception & Psychophysics 23, 459 (1978)
22. Kozlowski, L.T., Cutting, J.E.: Recognising the sex of a walker from a dynamic point-light display. Perception & Psychophysics 21, 575–580 (1977)
23. Runeson, S., Frykholm, G.: Kinematic specifications of dynamics as an informational basis for person-and-action perception: Expectation, gender recognition, and deceptive intention. Jounal of Experimental Psychology, General 112, 585–615 (1983)
24. Vallortigara, G., Regolin, L., Marconato, F.: Attraction to Motion. PLoS Biol. 3(7) (2005)
25. Polhemus, Liberty (2003) http://www.polhemus.com/
26. Bishop, C.: Neural Networks for Pattern Recognition. University Press, Oxford (1995)