Available online at www.sciencedirect.com
ScienceDirect
Cognitive Systems Research 7 (2006) 339–356
Cognitive Systems RESEARCH
www.elsevier.com/locate/cogsys

Human and machine perception of biological motion

Action editor: Risto Miikkulainen

Vijay Laxmi, R.I. Damper [*], J.N. Carter

Image, Speech and Intelligent Systems Research Group, School of Electronics and Computer Science, University of Southampton, Highfield, Southampton SO17 1BJ, UK

Received 5 February 2004; accepted 1 February 2006
Available online 19 April 2006

Abstract

More than 30 years ago, Johansson was the first to show that humans are capable of recovering information about the identity and activity of animate creatures rapidly and reliably from very sparse visual inputs – the phenomenon of *biological motion.* He filmed human actors in a dark setting with just a few strategic points on the body marked by lights – so-called moving light displays (MLDs). Subjects viewing the MLDs reported a vivid impression of moving human forms, and were even able to tell the activity in which the perceived humans were engaged. Subsequently, the phenomenon has been widely studied and many attempts have been made to model and to understand it. Typical questions that arise are: How precisely is the sparse low-level information integrated over space and time to produce the global percept, and how important is world knowledge (e.g., about animal form, locomotion, gravity, etc.)? In an attempt to answer such questions, we have implemented a machine-perception model of biological motion. If the computational model can replicate human data then it might offer clues as to how humans achieve the task. In particular, if it can do so with no or minimal world knowledge then this knowledge cannot be essential to the perception of biological motion. To provide human data for training and against which to assess the model, an extensive psychophysical experiment was undertaken in which 93 subjects were shown 12 categories of MLDs (e.g., normal, walking backwards, inverted, random dots, etc.) and were asked to indicate the presence or absence of natural human motion. Machine perception models were then trained on normal sequences as positive data and random sequences as negative data. Two models were used: a k-nearest neighbour (k-NN) classifier as an exemplar of 'lazy' learning and a back-propagation neural network as an exemplar of 'eager' learning. We find that the k-NN classifier is better able to model the human data but it still fails to represent aspects of knowledge about body shape (especially how relative joint positions change under rotation) that appear to be important to human judgements.

Keywords: Biological motion; Visual perception; Artificial perception; Cognitive modeling

1. Introduction

The ability to detect the characteristic motion of humans and other animals is a very important aspect of human vision. It remains largely unaffected by distance variations or poor visibility conditions. Even in poor quality videos or blurred images, humans not only perceive the motion as characteristic of a living being but can also discern the kind of activity, e.g., jumping, dancing, hopping, running or walking (Bobick & Davis, 1996; Cédras & Shah, 1995). Often we can recognise a friend walking at a distance. Any familiarity clues such as clothes, hair style etc. are mostly obliterated at large distances, so it must be the motion itself that is responsible for this identification. However, precisely what aspects of the visual scene are responsible for motion detection still remain unknown. Since human walking is a complex activity comprised of simpler movements – translatory and/or rotational, or more specifically pendular – a good understanding of gait perception may give insight into the visual perception mechanism generally (Stevenage, Nixon, & Vince, 1999). And from the computational point of view,

[*] Corresponding author. Tel.: +44 1073 594577; fax: +44 1073 594498.
E-mail address: rid@ecs.soton.ac.uk (R.I. Damper).

1389-0417/$ - see front matter © 2006 Elsevier B.V. All rights reserved.
doi:10.1016/j.cogsys.2006.02.002

a good understanding of the human perceptual system may lead to more robust computer vision systems with better noise tolerance and view-invariance as compared to existing ones.

Johansson (1973) was the first to show that humans are capable of recovering quickly and reliably information about the identity and activity of animate creatures from very sparse visual inputs. He filmed human actors in a dark setting with just a few strategic points on the body marked by lights – so-called moving light displays (MLDs). His experimental subjects reported a vivid impression of human movement, and were even able to tell the activity (e.g., walking, dancing, etc.) in which the actors were engaged. Only about 0.2 s was required for subjects to come to a judgement that the perceived patterns represented human motion. This is remarkable when we consider that only minimal, impoverished information is available in MLDs; each individual light point or 'dot' means little by itself and there are relatively few of them overall, yet somehow they are integrated over space and time to create a vivid and compelling global percept of the underlying motion. This has come to be called *biological motion* since similar reliable judgements can be made about moving light displays of non-human animal actors too. In this paper, however, we will focus entirely on human motion.

Subsequent to Johansson's pioneering work, the phenomenon of biological motion has been widely studied and many attempts have been made to model and to understand it. Several important questions arise: For instance, how exactly is the sparse low-level information integrated over space and time to produce the global percept? What is essential information and what is not? What transformations can MLDs tolerate while still maintaining the percept of biological motion? And how important is world knowledge (e.g., about human form, biomechanics of locomotion, forces of action and reaction, effects of gravity on bodies, and so on)? Although many of these questions can be (and have been) addressed by conventional, psychophysical experimentation, we believe that machine-perception models have a part to play. If a computational model can replicate human data then it might offer clues as to how humans achieve the task. Clearly, an artificial system is much easier to analyse and interrogate, so as to uncover its operating principles, than any living system could ever be. Further, if we can model human data using little or no world knowledge, then such knowledge cannot be essential to the perception of biological motion. Thus, there seems to be considerable potential in studying human and machine perception of biological motion in parallel. This paper, we believe, represents one of only few attempts to do so.

Although there do exist some works which attempt to model human perception of MLDs (e.g., Giese & Poggio, 2003; Goddard et al., 1992), these often make assumptions about the actual mechanisms of perception that we wished to avoid. (In particular, Giese and Poggio give an excellent account of possible neural mechanisms; The interested reader is recommended to consult this source for details.) As this is early work, we wanted to use unashamedly naïve models, based on general pattern recognition principles, omitting any biological detail. This avoids premature commitment to specific biological mechanisms as a basis of perception of MLDs that could only be provisional in the current state of knowledge. Admittedly, this also limits the possibilities for relating outcomes of our work to biology, yet we believe – with Dror and Gallogly (1999) – that there can be sound reasons for eschewing biological realism. As Dror and Gallogly point out (p. 173) "biologically implausible computational analyses can contribute to (1) understanding and characterising the problem that is being studied, (2) examining the availability of information and its representation, and (3) evaluating and understanding the neuronal solution". The remainder of this paper will, we believe, illustrate these contributions in the specific context of the perception of biological motion. Although, like us, Pollick, Lestou, Ryu, and Cho (2002) avoid commitment to biological details by using artificial neural networks, they address the different but related problem of gender recognition from MLDs. Similarly, Troje (2002) and Davies and Gao (2004) avoid building in biological details, by using principal component methods, but they too consider gender recognition.

The rest of this paper is structured as follows. Section 2 describes moving light displays in brief and gives an overview of human perception of such displays. Our methodological approach based on comparing human and machine perception is described in Section 3. Section 4 describes our video data of walking humans, and details how the various categories of MLD used in this study were derived from these data. We then describe in Section 5 details of the psychophysical experimentation from which we obtained human data both for training the machine-perception models and to act as a basis of comparison. Results of the human experimentation are presented in Section 6. Section 7 describes machine perception of MLDs. Section 8 discusses the results of the human and machine studies, and considers what we have learned from comparing them before concluding.

2. Human perception of biological motion: a brief review

Moving light displays (MLDs) can be obtained by affixing small lights to specific points of the object (a living being) and filming it in nearly dark conditions such that the resulting displays do not carry explicit information about the shape, structure or contours of the object. Another alternative is to attach reflector patches to the object and to film it in minimal lighting. In either case, the recorded film displays only the specific points. In perception studies related to human motion, these markers are usually attached to the major parts or joints of the body (e.g., head, shoulder, elbow, hip, knee, ankle, etc.) as shown in Fig. 1. For convenience, we will refer to all of

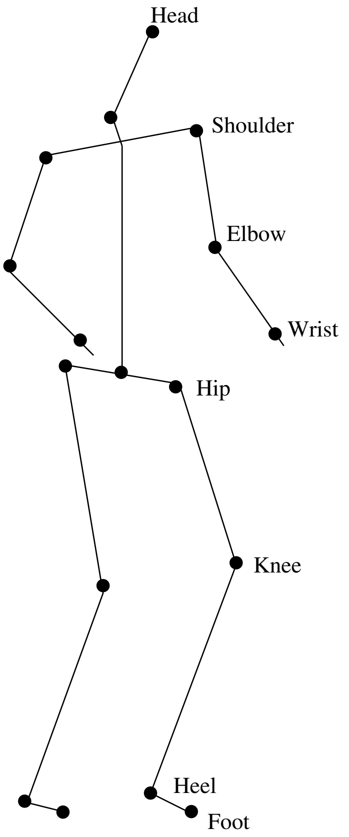


Fig. 1. Major joints of the human body. Only the joints used in this study have been labelled, and only those on the left side of the body. Note that head and foot are referred to as 'joints' for convenience.

these as 'joints' (although it is debatable that the head is a joint as such). In recent years, it has become technically possible to film the moving objects in full lighting and subsequently to mark the joints with video-editing software. The joints alone can then be displayed (e.g., on a computer screen) in a form of 'simulated' MLD. However they are derived, in respect of human perception, MLDs are minimal information systems as the motion and relative location of a small number of dots are the only available cues. In spite of this, the perception of human motion remains vivid.

Ever since Johansson (1973) used MLDs to demonstrate the capability of the human vision system to recover (or recreate) information about human form and activity, researchers have used them (or their computerised counterparts) extensively to study the mechanisms underlying visual motion perception. Johansson (1973, 1975, 1976) presented MLD sequences derived from humans carrying out various activities like walking, running, hopping, dancing and cycling to his experimental subjects. Twelve light markers were used (shoulder, elbow, wrist, hip, knee and ankle – both left and right). From the dynamic display, observers were not only able to discern a human figure but also to identify the type of motion correctly and without exception. They required a display time of approximately 0.2 s (about five frames) to achieve perceptual organisation of the pattern and to be able to make a correct report: "artificial" patterns with puppet-like motions required longer.

Subsequent work has shown that subjects are able to identify gender of actors from MLDs (Kozlowski & Cutting, 1977; Mather & Murdoch, 1994) and to tell individual friends (Cutting & Kozlowski, 1977). Even human infants as young as 3-months are able to perceive biological motion (Fox & McDaniel, 1982) as are non-human animals such as cats (Blake, 1993). Sumi (1984) found that when Johansson displays were inverted and run backwards, they were perceived more frequently as an upright image of a person moving forward in a very strange manner than as an inverted image of a person moving backward.

According to Ahlström, Blake, and Ahlström (1997), normal MLD sequences of walkers can be easily discriminated from phase scrambled ones. Phase scrambled sequences contain precisely the same local dot motions as regular MLDs but the starting point for the motion cycle of each is chosen randomly, e.g., the dot associated with the wrist may start at frame 6 whereas that associated with the elbow may start at frame 14. Dots specifying ankles only were perceived as non-biological motion but the addition of points representing knees resulted in an impression of biological motion. This impression grew stronger as more joints were added. Superimposition of an inverted figure on an otherwise normal biological motion sequence resulted in a multistable perceptual grouping of dots. However, this perceptual multistability was abolished when the dots describing the inverted figure were of different colour. Pavlova and Sokolov (2000) reported that despite prior familiarisation with an MLD figure at all orientations, its detectability within a mask (of distracting dots) decreased with a change in orientation from an upright figure (i.e., 0° reference) to one in the range 90–180°, where the latter is upside-down.

What then are the processes involved in perception of biological motion? Johansson (1975) suggested that the visual system follows the principles of central perspective and not Euclidean space and prefers maintaining figural constancy. Any change in the figure/shape is perceived as a central perspective transformation rather than the actual change happening in Euclidean space. Subsequently, Johansson (1976) formulated a mathematical model called "visual vector analysis" and suggested that the perception of biological motion involves an integration of spatio-temporal differentials, which can be abstracted as visual vector differentials. These differentials are, in effect, motion vectors of a joint. Thus, motion abstraction is akin to spatio-temporal differentiation. The visual memory needs to perform a continuous integration of these differentials for the grouping of perceptual elements.

Cutting and Proffitt (1981) suggested that while perceiving biological motion from MLD displays, the motions and locations of the hip and shoulder are extracted first as they are close to the body's deepest center of moment within the torso. These points now serve as the static centers of moment for the lower and upper body, respectively. For example, for the lower body, the hip acts as the center of moment for the knee, which moves in pendular fashion about it. One perceives the knee motion only by its motion relative to the hip.

Pinto and Shiffrar (1999) reported that detection of a figure missing elements ('joints') on the extremities, i.e., ankles and wrists as in Fig. 2(a), did not differ significantly from the detection of the whole figure. However, omission of central elements, i.e., hips and shoulders (Fig. 2(b)), did significantly diminish performance. Omission of mid-limb joints, i.e., knees and elbows (Fig. 2(c)), also impaired performance. On this basis, Pinto and Shiffrar concluded that a hierarchical vector analysis as suggested by Johansson (1973) and Cutting (1981) does not by itself give a complete description of the visual perception. The three body configurations missing extremities, mid-elements and central elements maintain a hierarchical structure amenable to such analysis yet sometimes performance is impaired. Hence, motion perception can not be explained by a simple hierarchical model. Rigid relations (cf. Hoffman & Flinchbaugh, 1982; Webb & Aggarwal, 1982) alone can not account for the different performances as all these configurations have the same number of rigid relations.

Another important finding of this study was that the detection of ipsilateral (arm and leg on same side) and/or contralateral limbs (both arms or both legs) did not differ significantly from that of diagonal limbs (arm and leg on opposite sides). On the basis of this observation, Pinto and Shiffrar concluded that neither dynamic symmetry nor the inclusion of information about principal axis is necessary for human motion detection. If dynamic symmetry was necessary, diagonal limbs would not be detected as human motion since these move in synchrony; further, no anti-phase information to indicate dynamic symmetry is available. If the elongated structure of the figure's principal axis was necessary, the contralateral limb condition would not be detected as only legs or arms were shown. Yet it is known that the absence of both, as in the case of randomly organised limbs (Fig. 3), largely abolishes the perception of biological motion.

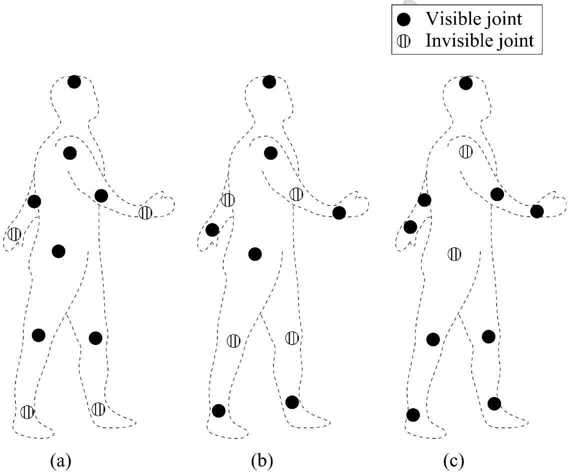


Fig. 2. Sub-configurations for the study of biological motion perception. In (a), extremities (wrists and ankles) are missing but performance is not significantly impaired. In (b), mid-limb elements (elbows and knees) are missing with a consequent loss of performance. In (c), central elements (shoulders and hips) are missing and again performance is impaired.

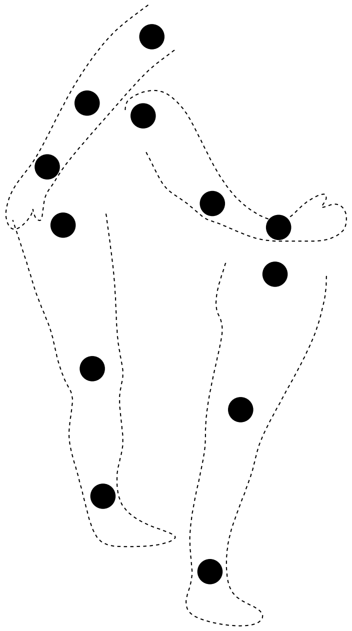


Fig. 3. Random ('spatial') scrambling of the limbs largely abolishes the detection of biological motion from MLDs.

Pinto and Shiffrar argued that the visual system responds equivalently to figures exhibiting any organisation of limbs consistent with the human form. Not only this, the visual system is capable of exploiting the configural information specifically indicative of the human form in the perception of MLDs. However, figural coherence is not sufficient to explain the detection of human movement. If it were so, the inverted figure detection would not significantly differ from the upright one. This difference, they argue, can not be explained by models based on hierarchical vector analysis and rigid relations. According to these authors, some other explanation is required.

3. Methodology: comparing human and machine perception

The primary objective of this work is to assess if it is possible to build an artificial perceptual system, making minimal assumptions, that is capable of reproducing the human data relating to biological motion detection. If so, this can act as a parsimonious model of human perception that is far easier to analyse and understand than the human system. It can also be used in future work to generate experimental hypotheses to be tested, so increasing our understanding of how humans are able to do biological motion detection.

As humans can perceive biological motion even from MLDs, where only spatio-temporal information of joint positions is available in the input, MLDs are used as the basis of this work. In particular, analysis of our machine models is likely to be very much simpler if it works with very sparse inputs, as in the case of MLDs, rather than the complete human shape being considered in full-frame video. To minimise the assumptions made in building the artificial perception model, we have adopted a machine-learning approach in which the model abstracts information about biological motion from (hopefully representative) example data with a minimum of intervention from us, the researchers. As machine learning is much more effective if it is *supervised*, i.e., the example (training) data are labelled with the correct classification, we have restricted our intervention to labeling data sequences as either *positive* (i.e., this is human motion) or *negative* (i.e., this is not human motion).

In an attempt to avoid our work being too sensitive to a particular choice of machine-learning methodology, we have deliberately used two very different approaches. Perhaps the most profound difference between methods is the distinction between *eager* and *lazy* learning (Aha, 1997). In eager learning, every effort is made (hence the name) to abstract and compress the training data into a small set of statistical regularities that capture the main generalisations of the domain. The archetypical eager learning methodology is error back-propagation (Rumelhart, Hinton, & Williams, 1986) where the training data are recoded into the connection weights and unit biases of an artificial neural network (ANN), which is then used as a data classifier or function interpolator. By contrast, lazy learning attempts to retain the training data in its entirety, often in the original form. Probably the 'purest' lazy learning technique is the k-nearest neighbour, or k-NN, method (Devijver & Kittler, 1982; Duda & Hart, 1973; Duda, Hart, & Stork, 2000). In the 1-nearest neighbour case, classification is effected by comparing an unknown instance with the entire database of examples, selecting the example that is 'closest' in some defined sense, and taking the label of this example to be the classification. More generally, the k-NN approach takes the classification to be that label that is maximally represented among the k nearest neighbours.

In our earlier work on machine perception of biological motion (Laxmi, Carter, & Damper, 2002a, 2002b), we attempted to train classifiers on different positive and negative examples of MLDS. In so doing, we tried to infer from the literature what might actually constitute positive and negative classes. This turned out to be problematic since the literature is not entirely clear and unambiguous on this issue. We therefore decided to collect our own human data to validate our labeling of the training data and to serve as a sound foundation for the comparison of machine and human perception of biological motion. In all, 12 different categories of MLD sequence were devised and used in the human experimentation and as input to the machine model, as we now describe.

4. MLD data and categories

In this section, we briefly describe the human motion data used as the basis for deriving MLDs. We also describe the various categories of MLD (e.g., normal, inverted, phase scrambled) that were used in subsequent studies of human and machine perception.

4.1. Dataset

The dataset used in this study was collected by Georgia Tech Research Institute (GTRI), Georgia Institute of Technology, Atlanta, GA. It consists of labelled sequences of 21 walkers. Joint labeling is three-dimensional and was achieved using infra-red markers. For each walker, there are four sequences:

(1) without shoes;
(2) without shoes and carrying a five-pound backpack;
(3) wearing street shoes; and
(4) wearing street shoes and carrying a five-pound backpack.

There is only one sequence for each walking mode; so, there are 84 sequences in total. Each walker has 15 infra-red markers ('joints'), namely head plus two each of shoulder, elbow, wrist, hip, knee, ankle, foot. We thus have three more markers for each walker than in the classical Johansson (1973) MLDs; the additions are head and two feet. Each sequence was 150 frames long (i.e., approximately 2.6 s as the inter-frame interval in the GTRI data is 17 ms). Because the subjects walked freely, each sequence starts in general at a different (random) point in the gait cycle.

For presentation to subjects on a computer screen, 2-D data are required. Particular views of the 3-D data were in most cases prepared by simply ignoring one of the three dimensions. Full details are given below. In one instance, however, all the 3-D data were used (to produce an oblique view).

4.2. Sequence categories

In all, 12 categories of 2-D MLD sequences were prepared; 11 of the 12 being derived from the GTRI 3-D dataset. From the perception studies as discussed in Section 2, MLDs of a walking or running human seen in side view are perceived as positive instances of human motion whereas a sequence of random configurations of dots is not (i.e., it is a negative instance). To determine if a machine is capable of perceiving human motion in a manner akin to humans, we need to determine what constitute positive and negative instances of human motion in the case of other views or transformations, i.e., other *categories* of MLD. The categories used here were chosen either because they have been used in human perceptual studies in the past, or because they seemed (to us) to raise interesting questions about biological motion.

NOR This category is the NORmal, fronto-parallel ('sideways') view of a person walking left-to-right. NOR sequences were produced from the 3-D dataset by ignoring the z-axis data. Although the labeling of this and all other categories is ultimately to be determined by the human experimentation described in Section 5 below, this category of sequence is reported as a positive instance in all previous perception studies.

DIR This change of DIRection category was generated by a spatial reflection of individual frames of the NOR sequences in the vertical plane and temporal reversal of the sequences to give right-to-left movement of the walker. We expect such sequences to be labelled as positive by subjects as biological motion detection is independent of the direction of motion (Johansson, 1976).

WBK This Walking BacKwards category was obtained by spatial reflection of individual frames of the NOR sequences but without temporal reversal. Hence, the figure appears to walk backwards in a left-to-right direction. Although human shape is preserved, it is relatively uncommon (although not unknown) to see people walking backwards. However, Johansson (1976) reports that sequences of this category invoke perception of human motion, so this category is expected to be positive.

INV This is an INVerted version of NOR, corresponding to a viewing angle of 180° (0° reference for NOR). Although inverted displays have relative spatial and motion relationships similar to normal ones, most structure-based perception theories cannot explain the negative response of human observers to this category.

TOP This is the view obtained from the TOP. It is produced from the 3-D dataset by ignoring the x-axis data. As there are no results in the literature on this category of MLD, we had no prior expectation as to how experimental subjects would label it.

OBQ Here viewing is at an OBliQue angle of 60°. This is an exceptional case in that we used all the 3-D data to produce the OBQ sequences. A simple program was written to extract these sequences from the 3-D data. Again, to the best of our knowledge this category has not previously been studied, so we had no prior expectation as to how experimental subjects would label it.

SPT A Small PerTurbation was added to all the joint positions in the NOR view. The perturbation was different from frame to frame. For any given frame, the bounding box is determined. Let w and h be, respectively, the width and height of this rectangle. Every joint, at position (x,y), is perturbed to a new position ($x + \mathrm{d}x$,$y + \mathrm{d}y$) according to the following pseudocode:

```
$d = 1.0/\mathrm{random}(6.0, 10.0),$
$\mathrm{d}x = x + \mathrm{random}(-d, d) * w,$
$\mathrm{d}y = y + \mathrm{random}(-d, d) * h,$
```

where random(a,b) returns a random number rectangularly distributed in the interval (a,b). In the worst case, a joint is perturbed by one-sixth of the dimensions of the body for that frame. This transformation retains the structure of the human figure on average, although inter-joint spatial relations are perturbed.

LPT A Large PerTurbation was added to all the joint positions in the NOR view. The perturbation was different from frame to frame. Perturbations are made to the joints as in the SPT case except that here the relevant pseudocode is

```
$d = 1.0/\mathrm{random}(3.5, 5.5),$
$\mathrm{d}x = x + \mathrm{random}(-d, d) * w,$
$\mathrm{d}y = y + \mathrm{random}(-d, d) * h.$
```

Thus, in the worst case, a joint is perturbed by approximately one-third of the dimensions of the body for that frame. This is of a sufficient magnitude that the human shape is significantly disordered and masked. Our expectation was that SPT would be labelled as positive but LPT would not.

PER The frames in a NOR sequence were PERmuted in a random order. Thus, each frame retains the human shape but the natural temporal order of walking is destroyed. It is expected that this will be labelled negative by subjects.

SSR Sequences in this category were derived by Spatially ScRambling the limbs – arms and legs – in the NOR view (cf. Fig. 3). Only the positions of the limbs are randomised; motion trajectories are not disturbed at all. These sequences were expected to be labelled negative in view of the results of Pinto and Shiffrar (1999).

PSR In this category, NOR sequences were Phase ScRambled as described in Section 2 above. These sequences were expected to be labelled negative in view of the results of Ahlström et al. (1997).

RAN The appropriate number of dots was placed at RANdom in the bounding box of each frame of the fronto-parallel (NOR) sequence. As this configuration has neither shape nor motion consistent with human walking, the sequence was expected to be labelled strongly negative by subjects. This is the only category of the 12 that was not derived from the GTRI 3-D dataset.

Sequences were scaled to have the same height for presentation to experimental subjects and to the machine-perception models. For each category, a snapshot of some frames from one of the corresponding image sequences is shown in Fig. 4.

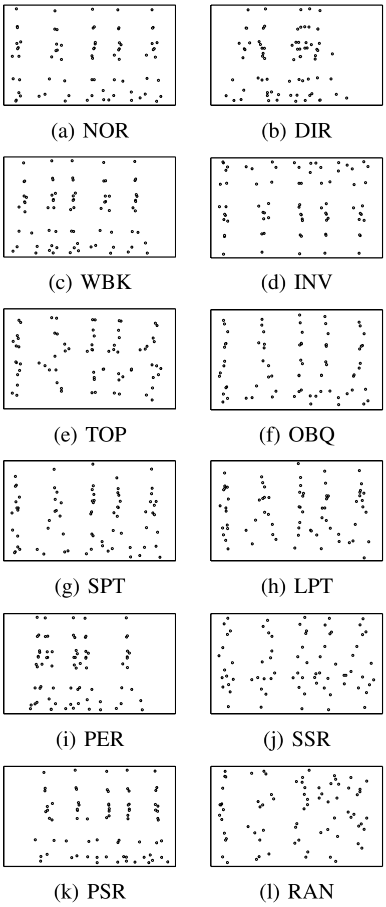
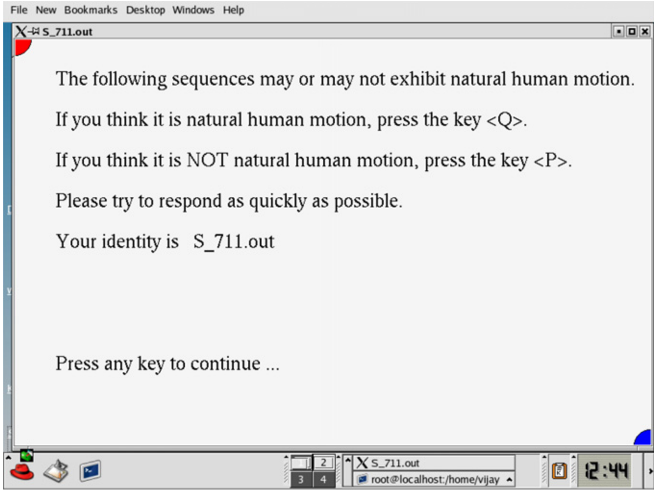

(a) NOR (b) DIR (c) WBK (d) INV (e) TOP (f) OBQ (g) SPT (h) LPT (i) PER (j) SSR (k) PSR (l) RAN

Fig. 4. Snapshots (of 5 frames) from each of the 12 MLD categories used in this study.

5. Details of human experimentation

To provide training data for the machine-learning models, and to give a basis of comparison between human and machine perception, an extensive psychophysical experiment was undertaken with 93 subjects. The subjects' task was to distinguish sequences displaying biological motion (positive examples) from sequences that did not (negative examples). As far as possible, we attempted to reduce any biasing of subjects to expect to see moving humans by carefully controlling their instructions and minimising what they were told about the experiment. Thus, they were not given any prior knowledge of the experimental set-up and were asked not to discuss it subsequently with any other subject. At the beginning of the experiment, they were given on-screen instructions as shown in Fig. 5. We also attempted to focus on basic perceptual classification (minimising post-perceptual reinterpretation of the data) by requiring subjects to respond as quickly as possible.

Of the 93 subjects, 37 were students or research staff from the University of Southampton, UK. The remaining 56 were undergraduate engineering students and faculty members from Malaviya National Institute of Technology, Jaipur, India. The age range was 18–35 years. A number of


File New Bookmarks Desktop Windows Help
S_711.out
S_711.out
root@localhost:/home/vijay
12:44

The following sequences may or may not exhibit natural human motion.

If you think it is natural human motion, press the key <Q>.

If you think it is NOT natural human motion, press the key <P>.

Please try to respond as quickly as possible.

Your identity is S_711.out

Press any key to continue ...

Fig. 5. Screen shot of instructions to experimental subjects.

the Southampton subjects worked in gait recognition specifically, or computer vision generally. We felt that these subjects were likely to produce a different pattern of results from subjects without this prior exposure; they were designated 'experts'. There were 23 such expert subjects, all from Southampton. Expert status was ascertained by a simple pre-experiment question: "Do you now, or have you ever, worked in computer vision or image processing?"

For each of the 84 sequences in the GTRI dataset (21 walkers, with and without shoes, with and without backpack, see Section 4.1), all 12 categories of MLD (as discussed in Section 4.2) were generated. As a result, the total number of sequences was 1008. Sequences were shown in four *conditions*:

Absolute. Dots corresponding to all body joints were shown with their corresponding translatory motion.

Spot. Translatory motion was removed. In the NOR case, this will give the impression of walking "on the spot".

Partial. One half of the dots (randomly selected from each individual frame) were shown.

Centroid. Only the centroid of the dots (with its corresponding translatory motion) was shown.

This gave a total of 4032 possible stimuli but this was too much data to present to subjects. So only a subset of 576 was actually used, constructed as follows. Of the 84 basic sequences described in Section 4.1, a selection of 12 was made by randomly taking 3 sequences from each of the 4 walking modes (with/without shoes, with/without backpack). This gave 12 selections times 12 categories (NOR, DIR, etc.) times 4 conditions (absolute, spot, etc.), or 576 sequences in all. To minimise the chance of results being biased by a statistically unrepresentative selection, each subject was shown a different random sample of 576 of the 4032 sequences.

The complete experiment was broken down into 6 sessions, during each of which 96 sequences were presented. Subjects were asked to view the screen from a comfortable distance of their own choosing. For the Southampton subjects, a 17-inch screen was used. For the Malaviya subjects, a 15-inch screen was used. Dot size was an ‘oval’ of 3 by 3 pixels centered on the screen coordinates of each joint. The original MLD data had a 60 Hz frame rate. We used a delay of 17 ms (closest integer value to 1000/60) between consecutive frames. After each 150-frame display, the screen was blanked for 300 ms. Subjects could take a short break after each session to help prevent fatigue.

For each sequence, subjects were required to press a key to indicate if it corresponded to natural human motion or not (cf. Fig. 5). If no key was pressed over the entire duration of a sequence, a time-out was recorded. For every subject, both the response itself and response time were recorded. For the purposes of analysis, a time-out was considered as a negative response.

6. Results of human experimentation

Here, we first deal with the rating responses for the absolute and spot conditions, which were generally the most interesting and informative. We then briefly discuss rating results for partial and centroid conditions, and finally present timing data.

6.1. Absolute and spot conditions

Fig. 6 summarises the positive responses obtained from all subjects (both expert and naïve) for the absolute and spot conditions. As each category was shown 12 times, a value of 6 positive responses (and 6 negative responses) corresponds to the chance rate as shown on the figure. Sample mean is shown as an asterisk and the error bars indicate the 95% confidence interval. Fig. 7 separates these results into those for the experts ((a) and (b)) and those for the naïve subjects ((c) and (d)), for absolute and spot conditions, respectively.

From these results, our task is to infer positive and negative labelings to inform our subsequent machine-learning study. For those categories where the mean was above the chance level of 6, a one-tailed *t*-test was used to determine with 95% confidence whether the category represents human motion (positive label; mean significantly greater than 6) or was indeterminate (mean not significantly different from 6). Similarly, for those categories where the mean was below chance, a one-tailed *t*-test was used to determine with 95% confidence whether the category was definitely not perceived as human motion (negative label; mean significantly less than 6) or was indeterminate (mean not significantly different from 6). In this way, we obtain the labelings shown in Table 1.

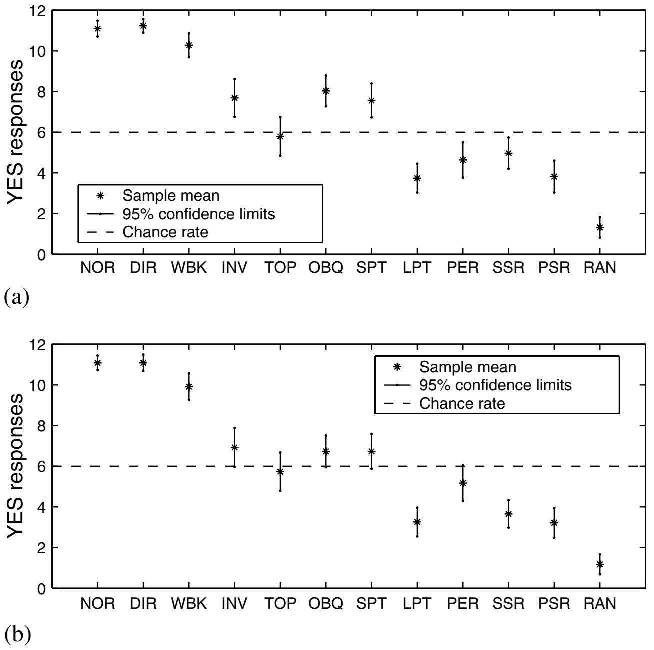


Fig. 6. Human responses for (a) absolute and (b) spot conditions. Asterisk denotes sample mean; error bars denote 95% confidence interval.

It is clear from Figs. 6 and 7 and Table 1 that the results for absolute and spot conditions are very similar. In fact, the main difference is that the spot responses are slightly more negative overall, probably reflecting the loss of information (about absolute motion) for this condition compared to the absolute condition. The similarity suggests that the human perceptual system is primarily sensitive to relative motion of the limbs between frames of MLDs rather than to absolute (translatory) motions of the figure across frames.

As expected, NOR, DIR and WBK are all labelled strongly positive by both expert and naïve subjects, while RAN is labelled strongly negative. However, the response for the upside-down display (INV) is not negative as expected. This contradicts the previous findings of Sumi (1984), Pinto and Shiffrar (1999) and Pavlova and Sokolov (2000). The direct comparison with the earlier experimental data is difficult since these previous workers used somewhat different stimuli (e.g., walkers in masking noise in the case of Pavlova and Sokolov) or experimental conditions (e.g., Sumi used totally naïve subjects and it may be that after a few trials our subjects were not naïve any more). An additional possibility that we favor is that, unlike the MLDs used by previous researchers, feet markers are displayed in our sequences and these act as additional cues to body-shape recognition for some subjects (more especially the experts).

The (novel) TOP category was found to be indeterminate, i.e., subject responses were not significantly different from chance. Yet OBQ is generally judged positive (an exception being the spot condition with the naïve subjects). It seems that the additional information that the oblique viewpoint offers about relative limb positions and movements can be exploited by subjects to build up an impression of human motion.

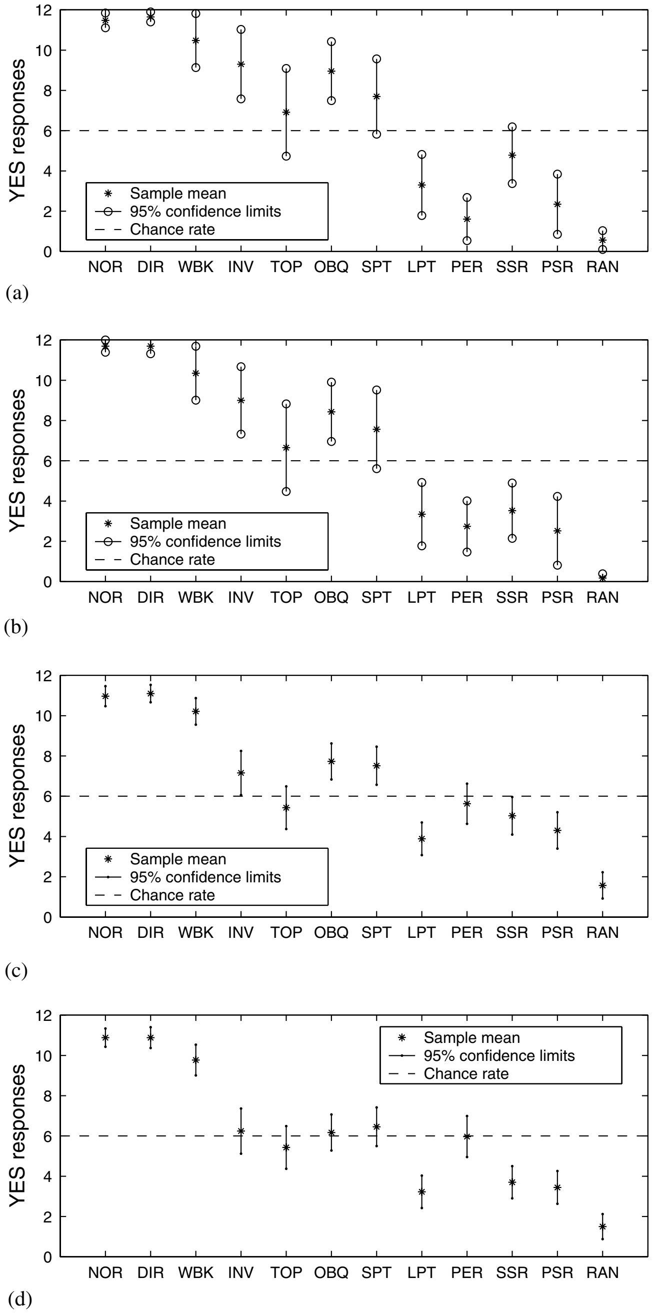


Fig. 7. Responses of experts for (a) absolute and (b) spot conditions and of naïve subjects for (c) absolute and (d) spot conditions. Asterisk denotes sample mean; error bars denote 95% confidence interval.

Consistent with our expectations, small permutations of the joint's spatial positions (SPT category) did not abolish the impression of human motion, although the positive responses were less strong. However, large perturbations (LPT category) led to consistent and moderately strong negative judgements.

Generally, the permuted, spatially scrambled and phase scrambled (PER, SSR and PSR) sequences were judged to be negative, in line with our expectations. The naïve subjects, however, gave markedly indeterminate responses to the PER category.

Table 1
Positive (+), negative (−) and indeterminate (blank) instances of biological motion as inferred from human data (93 subjects)

Category	Absolute			Spot		
	Expert	Naïve	Overall	Expert	Naïve	Overall
NOR	+	+	+	+	+	+
DIR	+	+	+	+	+	+
WBK	+	+	+	+	+	+
INV	+	+	+	+		+
TOP						
OBQ	+	+	+	+		+
SPT	+	+	+	+	+	+
LPT	−	−	−	−	−	−
PER	−		−	−		−
SSR	−	−	−	−	−	−
PSR	−	−	−	−	−	−
RAN	−	−	−	−	−	−

The naïve subjects are rather less assertive than the experts in that their mean responses are generally closer to chance, especially in spot condition. A category-wise analysis of variance indicates that responses for these two sets of subjects vary at the 5% significance level, as follows:

- INV, PER and PSR categories for absolute condition;
- NOR, INV, OBQ, PER and RAN categories for spot condition.

In all of these cases, the experts' responses are more categorical (i.e., further from chance). It is also noticeable from Fig. 6 that the variance of the experts' judgements is wider, but this is almost certainly a trivial consequence of there being fewer of them (23 as opposed to 70 naïve subjects).

6.2. Partial and centroid conditions

Fig. 8 shows consolidated human responses (i.e., expert, naïve and all subjects) for the partial and centroid conditions. In just the way that the spot results of the previous subsection display the same pattern as absolute results but shifted towards more negative responses, so the results for the partial condition are similar to those for spot condition but are again shifted downwards. This shows that the loss of information relative to absolute mode results in fewer positive responses (for spot and partial conditions) and further indicates that the partial condition – in which half of the joints are randomly masked on a per-frame basis – is more disruptive than the loss of absolute motion information in the spot condition. The centroid results are essentially uninteresting. All categories (even NOR) are rated negative. This is consistent with the inference, drawn from the similarity of results for the absolute and spot conditions, that translation between frames of the figure's centroid carries little or no information supporting positive identification of human biological motion.

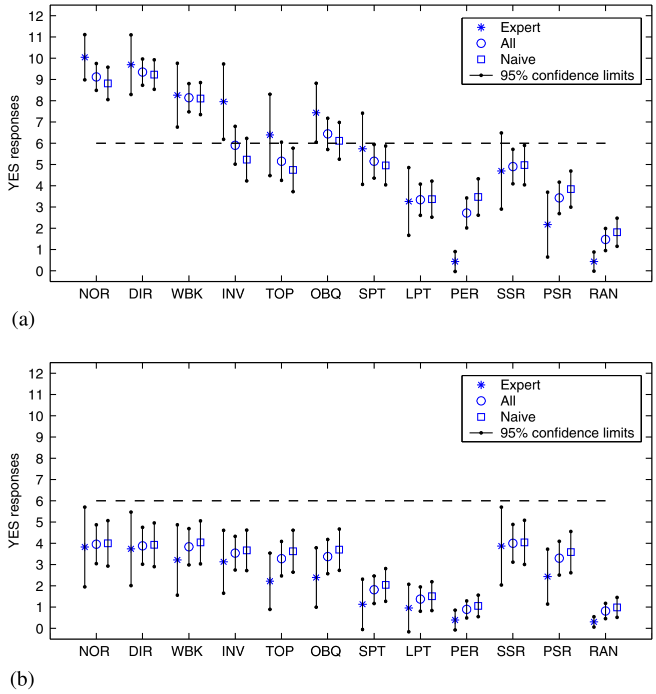


Fig. 8. Consolidated human responses for (a) partial and (b) centroid conditions.

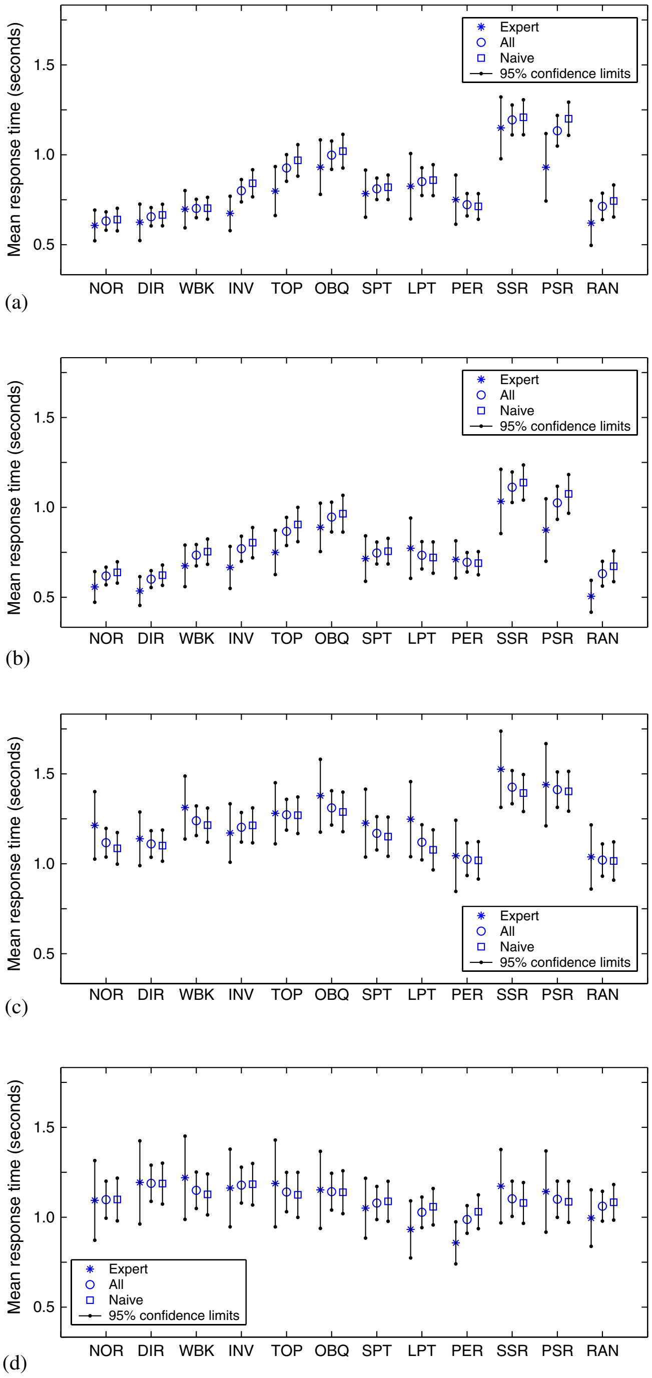


Fig. 9. Timing data for (a) absolute, (b) spot, (c) partial and (d) centroid conditions.

6.3. Timing data

Fig. 9 shows mean response times (excluding time outs) for the four conditions. The response is measured relative to the end of the sequence, and thus it does not necessarily indicate the time taken from the point at which the subject reached his/her decision to pressing a key to signal this decision. In particular, it could be negative. Nonetheless, conditions are constant for all subjects and all stimulus sequences, so we believe the data are both useful and meaningful.

Mean response times are very similar for absolute and spot conditions, with subjects generally taking longest to respond to SSR and PSR categories and responding fastest to NOR, DIR and RAN. The longest response times are seen with the partial condition, indicating that the rating task was hardest in this case. Again, SSR and PSR categories are the most difficult to rate, at least as assessed by response time, implying that the correct motion of each individual joint retained in these sequences was recognised by subjects but could not be integrated, spatially or temporally, respectively, into a realistic whole (hence the negative categorisation).

For the centroid condition, response times are relatively flat across categories and relatively long also, indicating that subjects saw little difference between categories (there was after all just a single dot) yet still were not able to make a speedy rejection of the sequences as instances of biological motion. It is difficult to know precisely why this is. Perhaps the residual information about human motion that is undoubtedly present (and that accounts for absolute condition judgements being overall slightly more positive than spot condition judgements) may have been recognised, but subjects were unwilling to rule that a single dot could be constitutive of human motion.

7. Machine perception of MLDs

Having obtained labelled human data for training and comparison purposes, we now consider the machine perception of biological motion from MLDs. As they were the most interesting and informative in the human data described in the previous section, we restrict attention to absolute and spot conditions only. (In any event, the centroid condition did not yield any positive human responses for any category, so there is nothing here to model.) The main points at issue are:

(1) Can a machine classifier be trained on data alone – without prior world knowledge or experience of physical laws and biological constraints governing movement of humans – to reproduce human responses to MLDs?
(2) If so, what can we learn about human perception from analysis of the trained classifier responses?

Since perception of biological motion is in essence a task of spatio-temporal integration, and most artificial pattern classifiers are basically recognisers of static patterns, our first requirement is to devise some means of presenting dynamic information to the k-NN and ANN models identified in Section 3.

7.1. Spatio-temporal integration for machine input

For any frame in an MLD sequence, the dot positions describe the body configuration at time t, where t is the position index of the frame in the sequence. The dots are scanned in a top-down, left–right 'raster' manner as shown in Fig. 10. It should be obvious that the correspondence of joints between frames will not be preserved in this scheme. That is, dot i in one frame will not necessarily correspond to the same joint as dot i in another frame (although it often will do so). The number of parameters per frame is twice the number of dots, as each dot is represented by two (x and y) coordinates. For this discussion, we will assume that an MLD sequence consists of F frames, with each frame containing M dots and, hence, $2M$ parameters/frame.

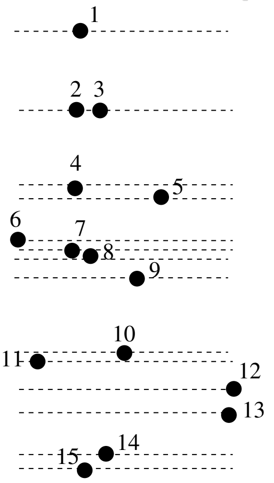


Fig. 10. Dots in an MLD sequence are scanned in a top-down, left–right ('raster') manner for input to the machine-perception model (k-NN or ANN). Numbers indicate the order in which the x–y coordinates are entered into the feature vector. In this example, there are $M = 15$ dots and, consequently, $2M = 30$ parameters.

Considerable thought was given to the suitability or otherwise for our purposes of this rasterised representation. Ideally, a bit-map with the same resolution as the computer display would have been used, but this implies enormous input dimensionality for an essentially sparse representation. Some sort of dimensionality reduction is essential from a practical point of view. We did not want to use a method such as principal components (e.g., Davies & Gao, 2004; Troje, 2002) as we felt that this transformation on the input space would complicate interpretation of the model. A particular consideration was that we wanted the machine-perception model to solve the problem of inferring correspondences between joints without undue 'help' from us, the experimenters, in just the way that human subjects have to do. From all these considerations, we felt that the raster-scan representation was a good compromise.

Using this representation, the body configuration vector $\boldsymbol{c}(t)$ corresponding to a single frame with index t is given by

$$\boldsymbol{c}(t) = (x_1(t), y_1(t), \ldots, x_M(t), y_M(t)) \quad t = 1, \ldots, F,$$

where $x_i(t)$ and $y_i(t)$ are the coordinates of the ith dot in the frame. Integration of temporal information over N frames can be effected by constructing an N-tuple $\langle \boldsymbol{c}(t), \boldsymbol{c}(t+1), \ldots, \boldsymbol{c}(t+N-1) \rangle$, i.e., a concatenation of the configuration vectors of N consecutive frames. We will refer to this concatenated vector as a *datapoint*. It is important to note that each datapoint is a sequence of static patterns; there is position information only and no explicit representation of motion in the form of velocity, for example.

For a given value of N, all possible datapoints, each with temporal information spanning N frames, are generated. So, for $N = 5$, the first datapoint was generated by concatenating configuration vectors for frames 1–5, the second by concatenating these vectors for frames 2–6, and so on. So, each datapoint is a snapshot of a fraction of a gait cycle. MLD sequences are presented to the classifier as an unordered collection of datapoints. As the snapshots overlap, this provides 'context' information, and is a classical way to transform data sequences into a form suitable for input to classifiers of static patterns (e.g., Sejnowski & Rosenberg, 1987).

The datapoints were normalised to the unit hypercube and then were subjected to mean removal. For the artificial neural network, such normalisation helps avoid network saturation and the mean removal helped faster convergence during training.

For a sequence consisting of F frames with M dots per frame and with groupings of N frames, the number of datapoints is $(F - N + 1)$ and the dimensionality of each datapoint is $2MN$. Fig. 11 illustrates the process of generating datapoints from an MLD sequence, with $N = 3$. There is a trade-off in setting the value of N: As this value increases,

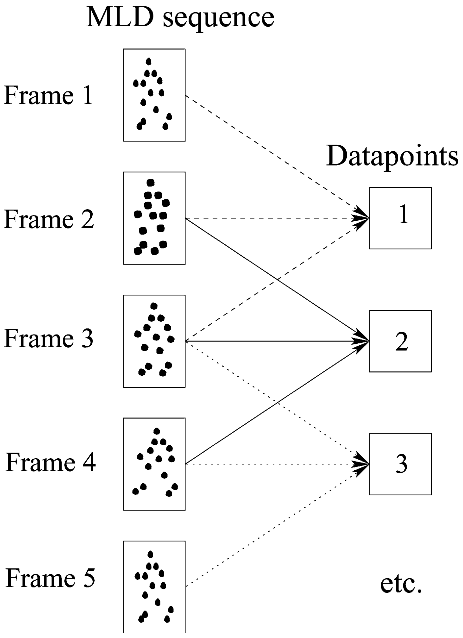


Fig. 11. Temporal integration of frames across an MLD sequence to generate datapoints (i.e., concatenated feature vectors for contiguous frames) in groups of $N = 3$.

the number of datapoints per sequence decreases (giving less training data, see below) but the temporal information available per datapoint increases. To explore the effect of this trade-off, we collected results for various values of N. Here, we will present results for $N = 1$, 9, 15 and 25.

The 12 categories considered in this work range from the well-structured (in terms of shape and motion) NOR category to the completely random RAN category. Human responses to these two categories were generally the most positive and most negative, respectively (Section 6). An exception is the DIR category, which was (not surprisingly) rated just as positive as NOR by the subjects. For all other categories, the responses are less differentiated.

7.2. Training and test data

For training the machine-perception systems, NOR sequences for 9 of the 21 walkers were taken as positive data; RAN sequences for the same 9 of the 21 walkers were taken as negative data. The test set consisted of the remaining datapoints. Thus, the training and test sets were mutually exclusive. Also, data were disjoint for the absolute and spot conditions. That is, when testing sequences in the absolute condition, the machine-perception models were trained using absolute data only; when testing sequences in the spot condition, they were trained using spot data only. There was one distinct model for each of $N = 1$, 9, 15 and 25.

7.3. k-Nearest neighbour detector

For the k-NN detector, each test datapoint is assigned the label of the nearest category in the respective training set. That is, $k = 1$ for the results reported in this paper. Different values of k were in fact studied, but no interesting dependence on k was found. Here, 'nearest' is defined in terms of Euclidean distance between corresponding dots in the datapoints. Generalisation results are presented in Fig. 12(a) and (b) for absolute and spot conditions, respectively. Each bar represents the fraction of test datapoints of the respective category labelled as positive biological motion.

In all cases, test (unseen) datapoints from the NOR category are classified as positive with 100% accuracy. Note that this is superior to average human performance but not dissimilar to the performance of several individual subjects (especially the experts). However, test data from the RAN category are classified as negative with considerably lower accuracy, even though this category constituted the negative training data. Nonetheless, there is a good degree of generalisation on the categories used in training.

It is noticeable that RAN test sequences are classified as more negative in spot condition than in absolute condition, and for the smaller values of N in constructing the datapoints. That is, in the absolute condition (Fig. 12(a)), the $N = 25$ model has greater difficulty in identifying negative instances than does the $N = 1$ model. This could either be because the absolute movement of the bounding box is incorrectly taken as an indicator of biological motion, or because of the mere presence of larger numbers of dots in approximately the right region – although actually in incorrect positions within that region. Looking to the RAN responses in the spot case (Fig. 12(b)), we see that this effect is much reduced, indicating that the movement of the bounding box (present in absolute but not in spot condition) is the principal explanation. Careful scrutiny of Fig. 12 reveals that the increase in positive responses as N increases is a quite general finding across categories, regardless of whether they are actually positive or negative. This points to an imperfection of the rasterised scan form of input, which seems not to be retaining motion information in a way which distinguishes positive and negative categories. Alternatively, it is possible that this is a real effect, which might be behind the positive votes for RAN in the human data (Figs. 6 and 7), which are more pronounced for naïve than for expert subjects.

Except for $N = 1$, where motion information is absent, the DIR category is labelled as 100% positive, as is the WBK category. Performance of the k-NN classifier for DIR sequences is typical of several of the human experts, but for WBK is slightly above that achieved by any of the human subjects. The result for DIR is intriguing. It shows either (1) that the machine-perception model has been able to generalise successfully from the left-to-right motion always present in the NOR sequences to unseen instances of right-to-left motion, or (2) such motion is being ignored. The latter possibility is consistent with the rasterised scan failing to represent this information appropriately. In either case, having the correct body disposition (as in the NOR, DIR and WBK categories but not RAN) is

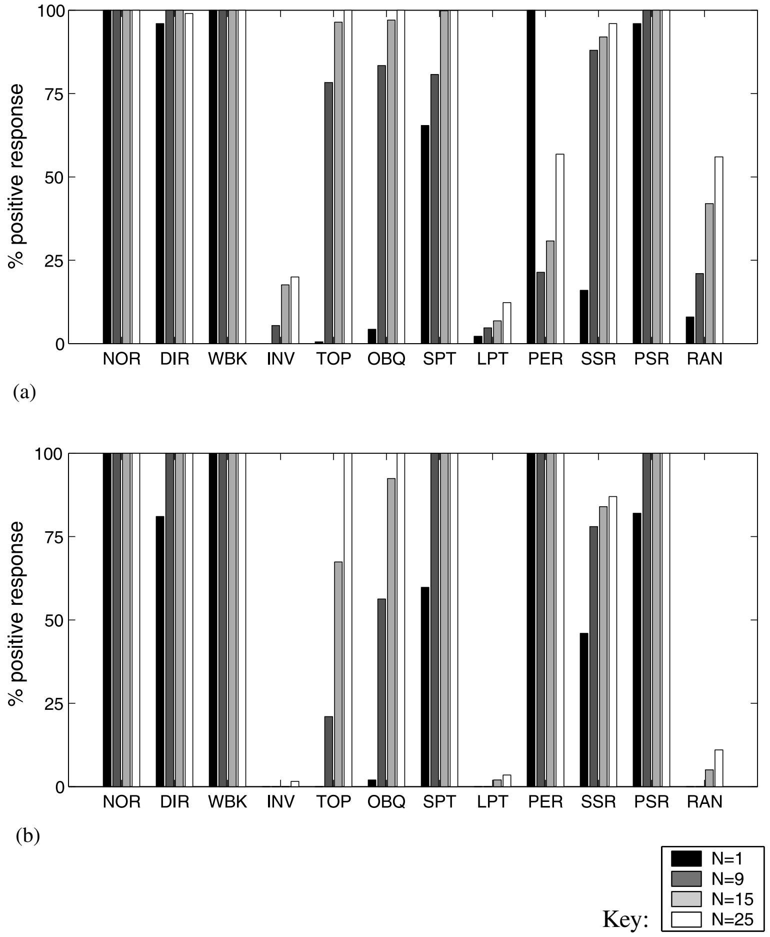


Fig. 12. Biological motion detection with *k*-NN model: generalisation performance on unseen test data in (a) absolute and (b) spot conditions. For each category, the performance is displayed for different values of frames per datapoint.

more important to the model than any absolute motion information.

Responses for INV are generally far more negative than the human data, where this category was weakly positive (Figs. 6 and 7 and Table 1). However, other investigators have generally found strong negative judgements in their human experiments. We conjectured earlier that the discrepancy between our human data and previous studies might be due to the inclusion of feet 'joints' in our MLDs, giving an additional cue to body-shape recognition. If correct, this kind of world knowledge (that human bodies have feet, two of them) would obviously not be available to the *k*-NN classifier. Again, like RAN, INV is more negative for the spot condition and for the smaller values of *N*. It seems that information about bounding-box motion is being used as the basis of detecting biological motion, whether appropriate or not.

Responses to TOP are interesting as for both absolute and spot conditions they range from strongly negative to strongly positive depending upon the value of *N*. It seems that provided enough context of the motion is seen, the *k*-NN classifier can discern it and treat it as indicative of biological motion. However, the evidence from other categories (e.g., RAN) is that the machine tends to use *any* motion (not just biological motion) as a cue to a positive judgement. And, of course, this category (TOP) is indeterminate for the human subjects.

In the human data, the shift of viewing angle from TOP to OBQ results in a shift of categorisation from indeterminate to positive (Table 1). This does not happen with the *k*-NN data, where the two categories are essentially indistinguishable. We take this as a strong indication that the human subjects can bring to bear prior knowledge about human body form (which is more discernible in OBQ than in TOP view) in a way that the machine cannot.

Like the human subjects, the *k*-NN classifier distinguishes SPT from LPT categories. However, the distinction is much stronger than that made by most human subjects. Indeed, the *k*-NN classifier is generally more categorical in its judgements, appearing to act somewhat as a 'super' expert (at least for those categories it gets right).

In the human data, the PER category was rated negative by experts and indeterminate by naïve subjects, irrespective of condition. By contrast, here the *k*-NN classifies PER sequences as indeterminate in absolute condition (more so for larger values of N) and strongly positive in spot condition. The 100% positive result for $N = 1$ is only to be expected since, in this case, the data are indistinguishable from NOR. The results for larger values of N indicate that relative joint position is a very strong positive cue for the machine, much stronger than is the movement of the joints in context. The latter, however, is not negligible since the enhanced movement information in the absolute condition is obviously contributing to the reduced positive ratings. And, of course, we have already seen in the case of RAN that *any* motion (represented statically in a datapoint of sufficient length) can be used as a positive cue. Overall, it seems that random motion has a centralising effect, pushing the data to midrange much as in the human data, but the effect is not as strong relative to body shape cues in machine perception as it is in human perception.

SSR and PSR categories for the k-NN classifier are quite positive, in contradiction of the negative human responses (Table 1). On the face of it, this is quite a puzzling finding as the correct cues to body shape are either absent (for SSR) or disordered (for PSR). However, the relatively high ratings found for $N = 1$ in each of the four cases of spot/absolute and SSR/PSR indicate that in spite of the scrambling some usable information about body shape is retained and the k-NN classifier is seemingly able to integrate this with whatever motion information is retained in the rasterised input.

In summary, the machine perception is more categorical than the human perception. This is perhaps not surprising as the k-NN classifier acts somewhat like a 'super' expert focusing solely on the task at hand and unaffected by fatigue, attentional issues, and/or other sources of inter- or intra-subject variability. Machine perception seems to be based primarily on static cues about relative joint position and only secondarily on motion cues, and these appear to be additive in their effects. The primacy of static cues is to be expected since the k-NN classifier is at heart a classifier of static patterns. Motion cues can only be extracted by integrating information quite widely across the feature space (Fig. 11) and it seems that the rasterised-scan input might not be retaining this information in usable form. Finally, there is some evidence from the INV and OBQ categories that human subjects do use prior knowledge of human body morphology to help make judgements, in a way which the machine classifier (because of its impoverished training and 'closed world') is unable to do.

7.4. Artificial neural network detector

A non-recurrent feed-forward network with back-propagation learning, a single hidden layer of two processing units ('neurons'), and a single output neuron was used. We decided against using a recurrent net, which might have been better suited to representation of the sequence information inherent in MLDs, because of known difficulties in training such nets and for comparability with the k-NN model. The units had sigmoidal activation functions. Initially, larger hidden layer sizes of 8 and 4 neurons were tried but results were insensitive to this so the smaller network was used subsequently. The learning rate and momentum for the back-propagation learning were 0.6 and 0.8, respectively. A positive/negative decision was made by simply thresholding the output neuron's activation at 0.5. Training was assumed to have converged if classification accuracy on the training set remained above 95% for five successive epochs after the assumed convergence point. Convergence was found to be very rapid: typically 5 or 6 epochs.

In the following discussion, the performance of the detector is averaged over five program runs, since back-propagation is known to be sensitive to initial condition (Kolen & Pollack, 1990). Weights were randomly initialised for each of these five runs. These initial weights were constrained in the range of $[-l,l]$, where $l = 1/2MN$, and $2MN$ is the dimensionality of vector input to a layer (Section 7.1).

Generalisation results are shown in Fig. 13(a) and (b) for absolute and spot conditions, respectively, for different values of N. Again, as in the case of the k-NN classifier, there is good generalisation on unseen examples of the NOR and RAN categories used in training. However, classification is much more categorical than humans, and even more categorical than the k-NN classifier. There is little evidence of any indeterminate categories at all.

Generally, the ANN takes NOR, DIR, WBK and SPT as positive categories as humans do, but unlike humans, PER and PSR are also taken as positive. An exception to this general observation (about the absence of indeterminacy) is the SPT category for $N = 1$ but, as there is no motion information retained in this case, this is probably just a reflection of the degree of perturbation applied in generating the SPT category. Interestingly, the ANN ratings are less positive for this specific situation (i.e., SPT, $N = 1$) than the k-NN ratings. This indicates that there is a loss of information in the 'eager' style of ANN training through compression relative to the 'lazy' style of k-NN training, which retains example data without compression. The 100% responses to PER (even for $N = 1$) show the ANN to be effectively insensitive to the motion information in the datapoints. INV, TOP, OBQ, LPT, SSR and RAN are all strongly negative for the ANN. This accords with the human data (Table 1) for the latter three categories but INV and OBQ should be positive whereas TOP should be indeterminate.

It seems that this simple ANN model is acting as a very crude dichotomiser and, as a result, it is unsatisfactory as a model of human perception. Perhaps its 'eager learning' compresses and/or discards too much. However, as mentioned above, an increase in hidden units does not improve the performance. Rapid convergence (see above) indicates that the machine can find discriminating features within the training set quite quickly, so the ANN is apparently doing something quite basic. It appears to be finding 'shortcuts' that are useful for distinguishing NOR from

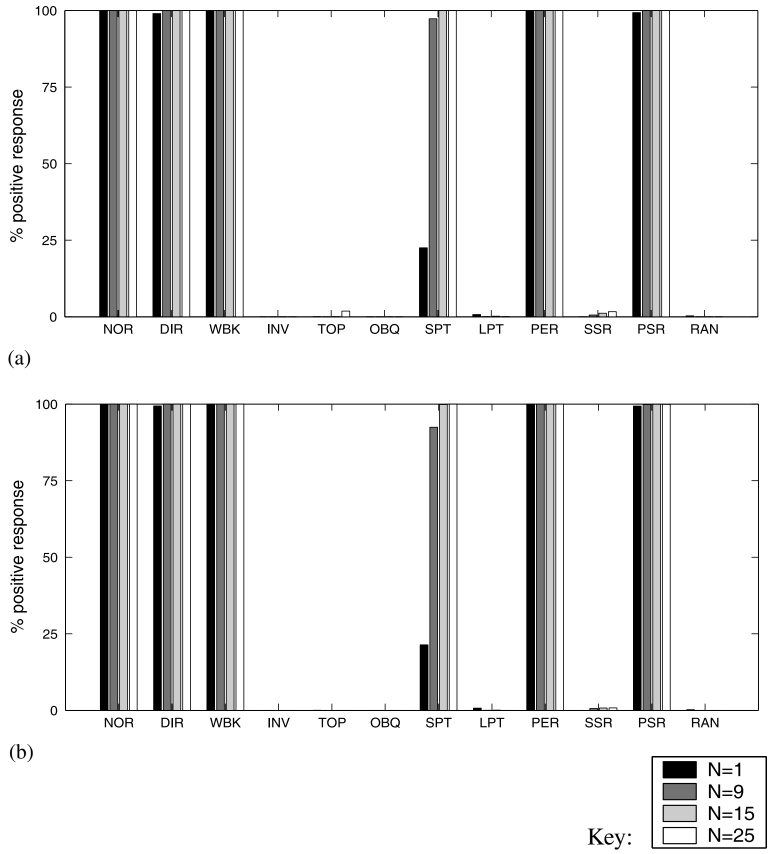


Fig. 13. Biological motion detection with ANN model: generalisation performance for (a) absolute condition and (b) spot condition.

RAN categories, but do not give good generalisation on some of the unseen categories.

One striking observation is that the spatial scrambling (SSR) produces strongly negative ratings whereas the phase scrambling (PSR) yields strongly positive ratings. The main difference between these is that the *x*–*y* coordinates of the joints are severely disrupted in SSR but less so in PSR. This led us to hypothesise that the ANN might simply (as a 'shortcut') be learning the absolute values and/or range of *x*–*y* positions occupied by the dots while remaining insensitive to the motion information encoded in the datapoints. This is consistent with what was found using the *k*-NN model, i.e., static cues appear to be primary since these simple models are essentially classifiers of static patterns and/or the rasterised-scan input is not retaining motion information in usable form.

If this hypothesis were true, then INV, TOP and OBQ would (as found) merit no positive responses as these are structurally different from the NOR category; the scanned dots occupy altogether different positions and hence different ranges. As the PER category has the same spatial structure as NOR but different temporal ordering, the hypothesis holds good. Also, for the PSR category, although different joints/dots have different initial phases, the range of spatial positions over the entire gait cycle remains the same.

To test this hypothesis, we retrained the ANN on data in which the height of individual MLD sequences was varied randomly in the range [0.2,1.0]. (Note that the height was fixed for all frames of any given sequence.) By disrupting the *y*-scale, this was intended to prevent the machine from focusing solely on spatial information, producing a simple dichotomy on this basis, at the expense of temporal information. Accordingly, our expectation is that the PER and PSR categories should, at least, become less strongly positive.

The responses of the retrained ANN are illustrated in Fig. 14. In contradiction of our hypothesis, these are broadly similar to those shown in Fig. 13 for the original ANN, i.e., the PER and PSR categories remain strongly positive in general. There are, however, two differences. First, there is a larger response for the LPT category, which is now close to being indeterminate. This is perhaps to be expected, since the random rescaling of the *y*-axis amounts to a large part of the large perturbation, making the test data more like the training data. Second, and very strikingly, positive responses to the $N = 1$ data are entirely obliterated in spot condition, but not in absolute condition. It seems that, in absolute condition with $N = 1$, the ANN is able to distinguish the NOR and RAN categories by some combination or correlation of the characteristic *x*-positions of the body and the relative limb positions in the former which are absent from the latter. In spot condition, this correlation is destroyed. For $N \geqslant 9$, however, the ANN is able to use the context information to infer *x*-motion characteristic of walking humans in the NOR case, without regard to any

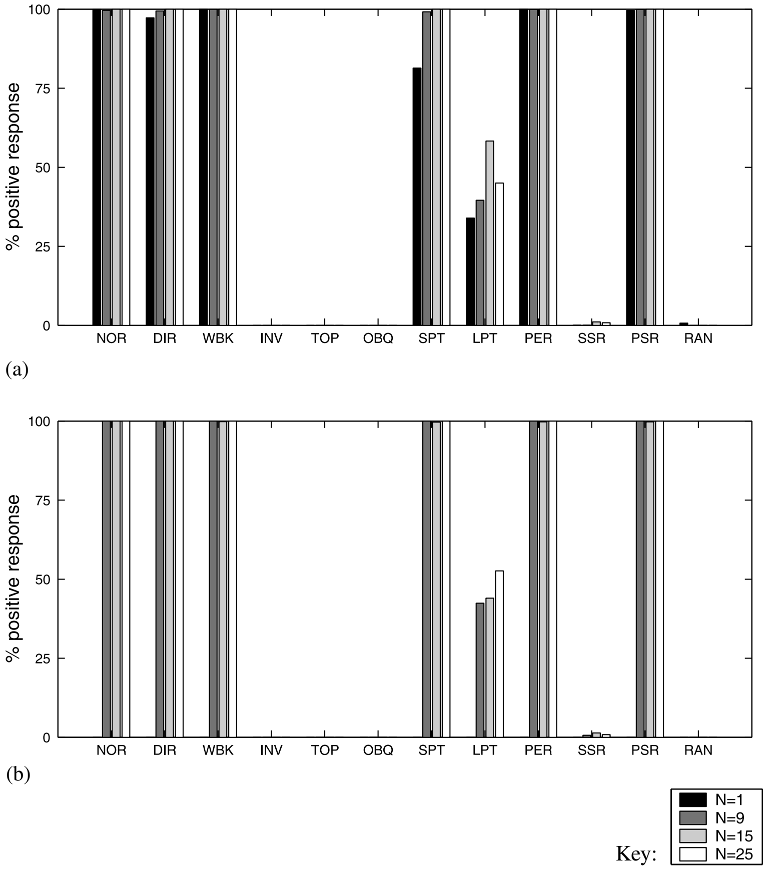


Fig. 14. Biological motion detection with ANN model trained on randomly resized data for (a) absolute condition and (b) spot condition.

correlation with relative limb positions, and to contrast this with the absence of such motion in the RAN case.

8. Discussion and conclusions

The main purposes of our psychophysical study of 93 subjects were to gain labeling data for training and assessing the machine-perception models but also to provide a basis of comparison with the machine-learning models. There was, however, one interesting observation from this study in its own right, namely that our subjects judge the INV category to be positive whereas the literature indicates this category to have been negative in previous studies. Although experimental differences between our study and earlier work are not precisely comparable, we have interpreted this in the light that the addition of feet to the MLDs gives our subjects additional information assisting body-shape recognition that was not available in earlier studies in which feet 'joints' were absent.

Regarding machine perception, the main questions addressed by our study were as follows (Section 7):

(1) Can a machine classifier be trained on data alone – without prior world knowledge or experience of physical laws and biological constraints governing movement of humans – to reproduce human responses to MLDs?
(2) If so, what can we learn about human perception from analysis of the trained classifier responses?

Results indicate that the answer to question (1) is essentially negative, at least as far as our attempt to do this is concerned. However, much will depend on the training, and on the way that the spatio-temporal information is represented and presented to the machine. The rasterised representation used here is only one of many possible and, as pointed out earlier, it only approximately maintains correspondence between joints across consecutive frames. At this stage, we know neither the extent to which correspondence is disrupted, nor the precise effects of this on results, although there is reason to think that our input representation has problems. A priority for future work is to explore the consequences of our input representation. One way that we might introduce motion information without "pre-solving" the correspondence problem in the way that we were reluctant to do in this work is to include 'difference images', much like the delta features used in speech and speaker recognition, where dots in each static image are replaced by vectors pointing to the nearest dot in the subsequent image.

Although we find that the *k*-NN classifier is a more accurate model of human performance than the ANN – perhaps because generalisation is less of an issue – neither is able to model correctly the (weakly) positive human responses to the INV category, and the differential human responses to the TOP and OBQ categories. We have conjectured that the human responses indicate that world knowledge (about body shape) is being used, and this sort of knowledge is only indirectly available to the machine-perception models through the NOR (positive) training data. As the machine-learning models do not have the capacity of rotation as humans do, a low response for the INV category is understandable. No doubt, responses more in line with the human data could be obtained by training the machine models on INV and OBQ data as positive instances and on TOP data as negative, but the strength of any such model must lie in its ability to generalise to unseen data. Ultimately, if we train on all 12 categories, any possible explanatory power is lost.

The largely positive responses for the PSR category for both *k*-NN and ANN models, indicate that the machine is less sensitive to the phase than to the static information. This almost certainly derives from the fact that these models are at heart recognisers of static patterns, as well as from imperfections of our input representation. Future work should consider models which are more capable of representing temporal, sequence information, such as hidden Markov models (HMMs). There is already some interest in the literature in the use of HMMs for gait analysis (Lee, Dalley, & Tieu, 2003; Meyer, 1997; Meyer, Pösl, & Niemann, 1998; Sundaresan, Roy Chowdhury, & Chellappa, 2003).

Consistent with our desire to minimise initial assumptions, the machine-learning methods used here are what are sometimes called parametric, or 'model-free'. That is, the learning system is given no 'higher-level' information about how limbs fit together to form bodies, how bodies move during walking, etc. other than that implicit in the training data. This is what we have characterised as 'world knowledge' throughout this paper. There is, of course, a long tradition in computer vision of model-based processing, including applications to human gait/walking (e.g., Hogg, 1983; Kale, Cuntoor, & Chellappa, 2002; O'Rourke & Badler, 1980; Rehg et al., 1995; Wachter et al., 1999; Yam, Nixon, & Carter, 2004). The indications of this work are that such approaches should be seriously considered in future work, as a way of breaking out of the 'closed world' assumptions of the *k*-NN and ANN classifiers.

Considering the answer to question 2, we cannot claim to have a machine-perception model that can be trained on data alone, without world knowledge, to replicate human data on detection of biological motion. On the contrary, as discussed above, we believe successful replication will require an approach in which world knowledge is supplied via some appropriate model of human body disposition and characteristic limb movements during walking. But this is not to say that the work reported here is in any way unsuccessful. In fact, we have learned a great deal, not least that human subjects appear to use prior knowledge to detect human motion in moving light displays, and a machine-perception model replicating the human data will need to do likewise. This points the way towards a realisation in the future of the potential that studying human and machine perception in parallel undoubtedly offers.

Acknowledgements

We gratefully acknowledge partial support by the European Research Office of the US Army under contract N68171-01-C-9002. Thanks are also due to Dr. Jon Geisheimer of Georgia Tech Research Institute for the use of the GTRI database.

References

Aha, D. W. (1997). Lazy learning. *Artificial Intelligence Review, 11*(1–5), 7–10.

Ahlström, V., Blake, R., & Ahlström, U. (1997). Perception of biological motion. *Perception, 26*(12), 1539–1548.

Blake, R. (1993). Cats perceive biological motion. *Psychological Science, 4*(1), 54–57.

Bobick, A. F., & Davis, J. W. (1996). An appearance-based representation of action. In *Proceedings of IEEE international conference on pattern recognition*, Vienna, Austria (pp. 307–312).

Cédras, C., & Shah, M. (1995). Motion-based recognition: a survey. *Image and Vision Computing, 13*(2), 129–155.

Cutting, J. E. (1981). Coding theory adapted to gait perception. *Journal of Experimental Psychology: Human Perception and Performance, 7*(1), 71–87.

Cutting, J. E., & Kozlowski, L. T. (1977). Recognizing friends by their walk: gait perception without familiarity cues. *Bulletin of the Psychonomic Society, 9*(5), 353–356.

Cutting, J. E., & Proffitt, D. R. (1981). Gait perception as an example of how we may perceive events. In R. Walk & H. L. Pick (Eds.), *Intersensory perception and sensory integration* (pp. 249–273). New York, NY: Plenum.

Davies, J. W., & Gao, H. (2004). An expressive three-mode principal components model for gender recognition. *Journal of Vision, 4*(5), 362–377.

Devijver, P. A., & Kittler, J. (1982). *Pattern recognition: A statistical approach*. Englewood Cliffs, NJ: Prentice-Hall.

Dror, I. E., & Gallogly, D. P. (1999). Computational analyses in cognitive neuro-science: in defense of biological implausibility. *Psychonomic Bulletin and Review, 6*(2), 173–182.

Duda, R. O., & Hart, P. E. (1973). *Pattern classification and scene analysis*. New York, NY: Wiley.

Duda, R. O., Hart, P. E., & Stork, D. G. (2000). *Pattern classification* (2nd ed.). New York, NY: Wiley.

Fox, R., & McDaniel, C. (1982). The perception of biological motion by human infants. *Science, 218*(4571), 486–487.

Giese, M. A., & Poggio, T. (2003). Neural mechanisms for the recognition of biological movements. *Nature Reviews Neuroscience, 4*(3), 179–192.

Goddard, N. H. (1992). The perception of articulated motion: recognizing moving light displays. Ph.D. Thesis, Department of Computer Science, University of Rochester, Rochester, NY.

Hoffman, D. D., & Flinchbaugh, B. E. (1982). The interpretation of biological motion. *Biological Cybernetics, 42*(3), 195–204.

Hogg, D. C. (1983). Model-based vision: a program to see a walking person. *Image and Vision Computing, 1*(1), 5–20.

Johansson, G. (1973). Visual perception of biological motion and a model for its analysis. *Perception and Psychophysics, 14*(2), 201–211.

Johansson, G. (1975). Visual motion perception. *Scientific American, 232*(6), 76–89.

Johansson, G. (1976). Spatio-temporal differentiation and integration in visual motion perception. *Psychological Research, 38*, 379–393.

Kale, A., Cuntoor, N., & Chellappa, R. (2002). A framework for activity-specific human identification. In *Proceedings of IEEE international conference on acoustics, speech, and signal processing (ICASSP '02)* Orlando, FL (Vol. 4, pp. IV-3660–IV-3663).

Kolen, J., & Pollack, J. B. (1990). Back-propagation is sensitive to initial conditions. *Complex Systems, 4*(3), 269–280.

Kozlowski, L. T., & Cutting, J. E. (1977). Recognizing the sex of a walker from a dynamic point-light display. *Perception and Psychophysics, 21*(6), 575–580.

Laxmi, V., Carter, J. N., & Damper, R. I. (2002a). Biologically-inspired human motion detection. In *10th European symposium on artificial neural networks (ESANN '2002)*, Bruges, Belgium (pp. 95–100).

Laxmi, V., Carter, J. N., & Damper, R. I. (2002b). Biologically-inspired human gait classifiers. In *Workshop on automatic identification advanced technologies (AutoID'02)*, Tarrytown, NY (pp. 17–22).

Lee, L., Dalley, G., & Tieu, K. (2003). Learning pedestrian models for silhouette refinement. In *Proceedings of IEEE international conference on computer vision*, Nice, France (pp. 663–670).

Mather, G., & Murdoch, L. (1994). Gender discrimination in biological motion displays based on dynamic cues. *Proceedings of the Royal Society of London (B), 258*(1353), 273–279.

Meyer, D. (1997). Human gait classification based on hidden Markov models. In *Proceedings of 3D image analysis and synthesis*, Erlangen, Germany (pp. 139–146).

Meyer, D., Pösl, J., & Niemann, H. (1998). Gait classification with HMMs for trajectories of body parts extracted by mixture densities. In *Proceedings of British machine vision conference (BMVC'98)*, Southampton, UK (pp. 459–468).

O'Rourke, J., & Badler, N. I. (1980). Model based image analysis of human motion using constraint propagation. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 2*(6), 522–536.

Pavlova, M., & Sokolov, A. (2000). Orientation specificity in biological motion perception. *Perception and Psychophysics, 62*(5), 889–899.

Pinto, J., & Shiffrar, M. (1999). Subconfigurations of the human form in the perception of biological motion displays. *Acta Psychologica, 102*(2–3), 293–318.

Pollick, F. E., Lestou, V., Ryu, J., & Cho, S.-B. (2002). Estimating the efficiency of recognizing gender and affect from biological motion. *Vision Research, 42*(20), 2345–2355.

Rehg, J. M., & Kanade, T. (1995). Model-based tracking of self-occluding articulated objects. In *Proceedings of international conference on computer vision*, Cambridge, MA (pp. 612–617).

Rumelhart, D. E., Hinton, G. E., & Williams, R. (1986). Learning representations by back-propagating errors. *Nature, 323*(9), 533–536.

Sejnowski, T. J., & Rosenberg, C. R. (1987). Parallel networks that learn to pronounce English text. *Complex Systems, 1*(1), 145–168.

Stevenage, S. E., Nixon, M. S., & Vince, K. (1999). Visual analysis of gait as a cue to identity. *Applied Cognitive Psychology, 13*(6), 513–526.

Sumi, S. (1984). Upside-down presentation of the Johansson moving light-spot pattern. *Perception, 13*(3), 283–286.

Sundaresan, A., Roy Chowdhury, A., & Chellappa, R. (2003). A hidden Markov model based framework for recognition of humans from gait sequences. In *Proceedings of IEEE international conference on image processing*, Barcelona, Spain (Vol. 2, pp. 85–88).

Troje, N. F. (2002). Decomposing biological motion: a framework for analysis and synthesis of human gait patterns. *Journal of Vision, 2*(5), 371–387.

Wachter, S., & Nagel, H.-H. (1999). Tracking of persons in monocular image sequences. *Computer Vision and Image Understanding, 74*(3), 174–192.

Webb, J. A., & Aggarwal, J. K. (1982). Structure from motion of rigid and jointed objects. *Artificial Intelligence, 19*(1), 107–130.

Yam, C. Y., Nixon, M. S., & Carter, J. N. (2004). Automatic person recognition by walking and running via model-based approaches. *Pattern Recognition, 37*(5), 1057–1072.