Biol. Cybern. 91, 306–314 (2004)
DOI 10.1007/s00422-004-0516-0

Biological Cybernetics

A stochastic model for the detection of coherent motion

Jason Lee[1], **Willy Wong**[1,2]

[1] Edward S. Rogers Department of Electrical and Computer Engineering, University of Toronto, 10 King's College Road, Toronto, ON, M5S 3G4, Canada
[2] Institute of Biomaterials and Biomedical Engineering, University of Toronto, 4 Taddle Creek Road, Toronto, ON, M5S 3G9, Canada

Received: 4 March 2004 / Accepted: 5 August 2004 / Published online: 8 October 2004

Abstract. A computational model is presented for the detection of coherent motion based on template matching and hidden Markov models. The premise of this approach is that the growth in detection sensitivity is greater for coherent motion of structured forms than for random coherent motion. In this preliminary study, a recent experiment was simulated with the model and the results are shown to be in agreement with the above premise. This model can be extended to be part of a more complex and elaborate computational visual system.

1 Introduction

In the 1970s, Johansson (1973) demonstrated that the human visual system is remarkably sensitive to motion of a human or biological origin. The movements of the joints alone were sufficient to convey full information about the gender of a person (Kozlowski and Cutting 1977; Mather and Murdoch 1994) and his/her activity (Dittrich 1993; Brownlow et al. 1997). An example of a Johansson representation or display is shown at the top of Fig. 1. The joints of the body alone are illuminated, but the remainder of the human form is not visible. A recent study by Neri et al. (1998) further demonstrated that for human observers the growth in detectability of biological motion is greater than for simple translational motion in a random dot kinematogram (see bottom of Fig. 1). While biological and translational motion are both examples of coherent motion, they differ in terms of complexity and detectability. Naively one might expect that the simpler motion type (i.e., translational motion) is easier to detect because of the greater redundancy in the motion sequence.

In the study by Neri et al. (1998) detection and discrimination thresholds were estimated from biological motion with human observers. Biological motion was derived from a Johansson display showing the temporal sequence of illuminated points of a moving person. A Johansson display can be generated by recording either the motion of a live subject or by using computer algorithms like the one detailed in Cutting (1978). The walking figure or signal was masked by the inclusion of dots that are uncorrelated to the signal. The amount of signal was varied by displaying from 1 up to a maximum of 11 light sources (top of Fig. 2). The subject's task was to correctly identify the presence or absence of a walking person in a single-interval forced-choice task. The number of noise dots was then varied until a certain performance criterion was achieved (e.g., a 75% correct response). This measure of performance therefore yielded an estimate of the absolute sensitivity or threshold of detecting biological motion. These results were then compared to a different experiment with translational motion in random dot kinematograms. Added noise was again varied as a function of signal, while performance level was kept constant. In both cases, it was found that the signal and the noise obeyed a relationship of the form

$$N \propto S^{\alpha} , \tag{1}$$

where α took on values greater than 1 for biological motion and 1 for translational motion. Equation (1) suggests that the increase in sensitivity for biological motion is greater than that of random coherent motion for the following reason. In the case where the exponent is, say, 2, a twofold increase in the signal will result in a fourfold increase in tolerable noise. Conversely, when α equals 1, the tolerable noise simply increases linearly with the signal. We feel that (1) is a result of critical importance and have built a preliminary model of motion detection around this result. While there have been models proposed for the detection of biological motion (e.g., Cutting and Proffitt 1981; Goddard 1992; Troje 2002), no known computational models currently address (1).

A basis for understanding (1) comes from the analysis of the underlying statistics of the system (e.g., Tripathy et al. 1999). In the case of biological motion, detectability is limited by the variability of noise from trial to trial. Since the masking noise is Poisson distributed, we know that the variance is proportional to the number of noise

Correspondence to: W. Wong
(e-mail: willy@eecg.utoronto.ca
Tel.: +1-416-9788734, Fax: +1-416-9468734)

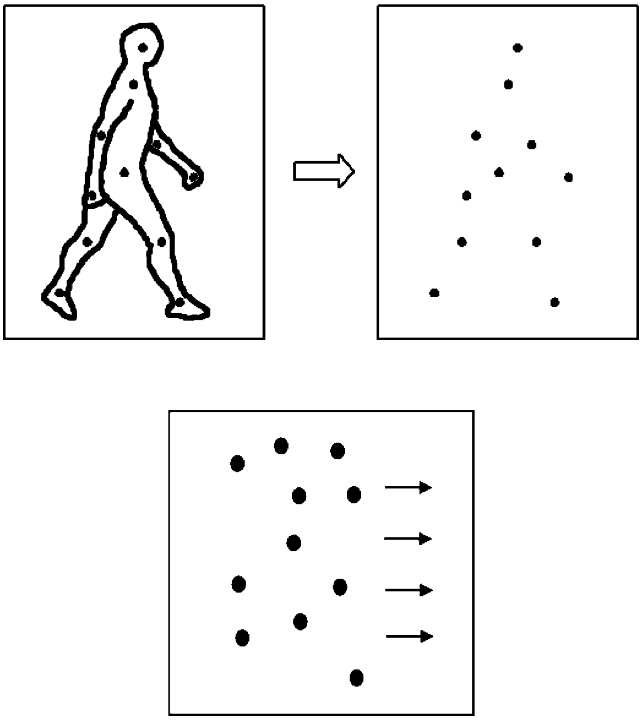

Fig. 1. *Top*: A Johansson display illustrated for a walking person. Only the joints of the person are illuminated and nothing else is displayed. The net dot displacement over the entire figure is equal to zero as the figure always remains centered in the display. *Bottom*: Translational motion in a dot kinematogram. The *dots* undergo horizontal displacements of equal steps in successive frames, and hence the net displacement here is nonzero. Image sizes are not to scale

dots (Wong and Barlow 2000). For constant detectability, $d' = \Delta\mu/\sigma =$ constant, where $\Delta\mu$ is the amount of signal S and σ is proportional to the square root of the noise N. From here we obtain $N \propto S^2$. In the case of translational motion detection, correspondence noise becomes the dominating factor and the number of spurious pairs increases with the square of the number of noise dots (e.g., Barlow and Tripathy 1997). Hence the standard deviation grows linearly with the noise N. Following a similar argument we obtain $N \propto S$. As we shall see, the computational model presented in this paper captures these results.

2 Background and motivation

There have been extensive studies into the theories of motion perception and in the development of computational models of the visual system (e.g., Marr 1982; Adelson and Bergen 1985; Dawson 1991; Goddard 1992; Fredericksen et al. 1993; Grzywacz et al. 1995). However, many of these models are either centered around the detailed neural architecture of the visual processes (e.g., Goddard 1992; Giese and Poggio 2003) or are largely computational in nature (e.g., Dawson 1991; Fredericksen et al. 1993). Fewer studies have addressed high-level, systems-oriented models. We propose one such model based upon the notion of signal detectability.

It is widely recognized that signal detection theory is the standard tool used in the analysis of threshold phenomena (e.g., Green and Swets 1966). Central to the signal-detection approach is the idea that the internal representation of a sensory signal is a random variable determined from the responses of neurons involved in the processing of sensory information. While little is known about the nature of the internal response in motion processing, studies have identified the neurons from the areas MT (V5) and MST to be of importance in random dot kinematograms (Britten et al. 1992; Newsome et al. 1989; Celebrini and Newsome 1994; Braddick et al. 2000) and the region of the superior temporal sulcus (STS) in the perception of biological motion (Grossman et al. 2000; Vaina et al. 2001).

We have taken a different route here and have largely ignored the transduction process from image to neural response. Instead we focus on the stochastic dynamics associated with the internal mental representation of the sensory signal. We have chosen a hidden Markov process to model the temporal variability associated with the internal representation. The motivation of this choice is now discussed.

Hidden Markov modeling is a probabilistic technique for the study of time series. Hidden Markov processes differ from Markov processes in that there are additional mechanisms that alter the observable sequence. Figure 3 illustrates the differences between the two models. Shown at the top is a Markov series in which each state (gray circle) stochastically determines (horizontal arrow) the next state. It is well known that most real phenomena are not well described by such a process. The lower half of Fig. 3 is an example of a hidden Markov model (HMM). The Markov chain is still present (gray circles), but the observable outcomes (white circles) are determined by additional stochastic steps (vertical arrows). In HMM terminology, the gray nodes are the "hidden states" while the white nodes are the "observable states." Other architectures are possible (e.g., additional layers of hidden states), although they will not be considered in this paper. Those who are unfamiliar with HMMs are invited to consult a number excellent references (e.g., Poritz 1988; Rabiner 1989).

A HMM can be used to detect temporal patterns under noisy conditions. The model is first trained with a noise-free target pattern. Given an input, the output of the model is a likelihood representing the probability that the input was produced by the same model that generated the training pattern. If the two patterns are highly correlated, this likelihood will be high. Two problems now emerge in connection with the use of HMMs in our model. The first problem is to associate the likelihood or output of the HMM to a decision variable in the detection model, and the second problem is to devise a technique whereby the HMM can process the input signal. The solution to the first problem is straightforward: we simply take the likelihood of the HMM to be the decision variable. If the likelihood exceeds a predetermined threshold, the input sequence is deemed to be generated by the same source as the training sequence. Regarding the second problem, recall that the Johansson display is a sequence of frames encoding the motion of a moving person. Several of these frames were selected initially as templates for the decision model. The input was matched to the templates, and a sequence of numbers was generated representing the number

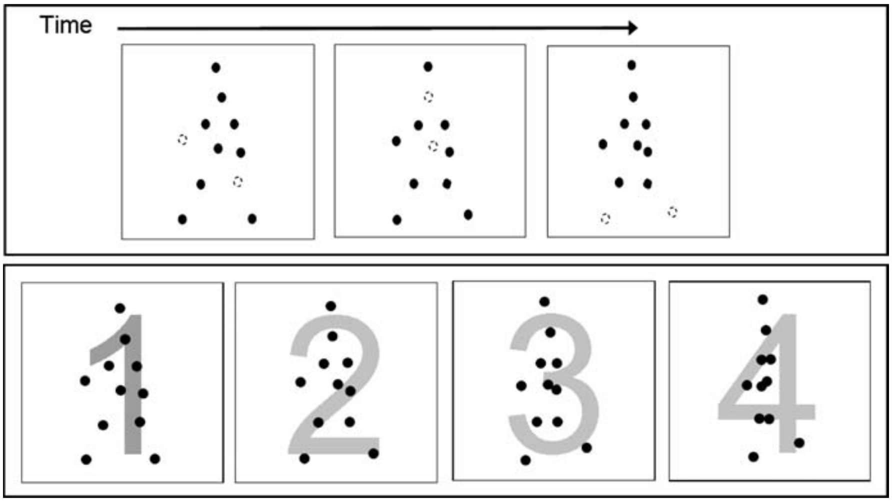

Fig. 2. *Top*: The amount of signal displayed is controlled by randomly selecting *dots* in each frame. An example is shown here where 9 *dots* out of 11 are selected in each frame. *Bottom*: A total of four templates were chosen at regular intervals from the full walking sequence. These templates were then used in the template-matching process. The walking sequence itself was generated following the algorithm by Cutting (1978)

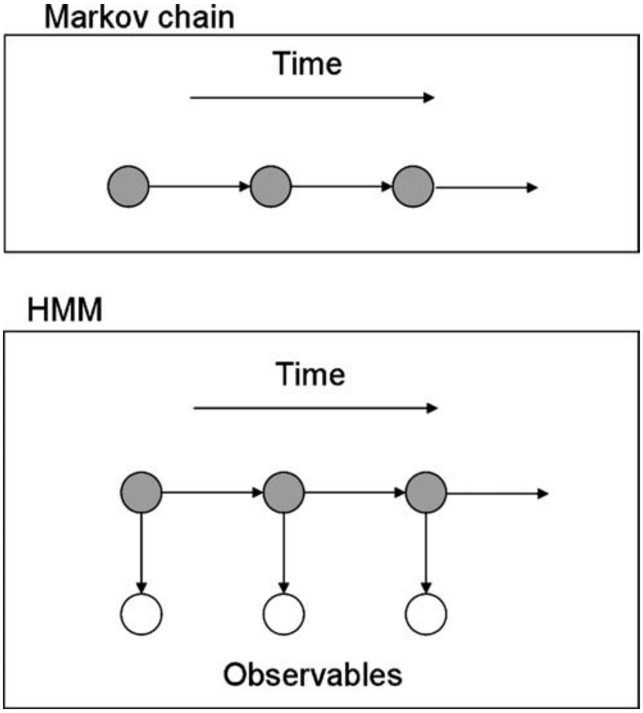

Fig. 3. Schematic diagram illustrating the difference between a Markov process and a hidden Markov model. In a Markov process, the current state depends stochastically on the previous state. In an HMM, the Markov process is still present (*gray circles*), but there are additional transitions (*vertical arrows*) that affect the outcome

of the template that corresponded to the best match. As an example, consider the case where four frames were chosen from a walking person sequence (see bottom of Fig. 2). Applying the templates (numbered 1–4) to a noise-free walking signal yields an output sequence like “1 1 1 2 2 2 3 3 3 4 4 4,” etc. On the other hand, if the same templates were applied to a sequence uncorrelated to the moving walker, the resulting sequence is random, e.g., “2 4 3 1 2 3 1 2 2 3 4 3,” etc. These output sequences are processed by the HMM and a likelihood is generated.

The process described above is similar to the idea of matched-filter analysis. The premise is that patterns are learned and stored in memory throughout life. When the sensory system engages in pattern recognition, the observed patterns are matched to internal templates and the closest match becomes the object “perceived.” The idea of matched-filter is not entirely unreasonable, and evidence for a biological basis has been found in many studies (e.g., Peterson and Gibson 1991; Wong and Barlow 2000).

In our simulations, we have introduced additive masking noise that is uncorrelated with the signal. The noise has a limited lifetime of one frame. The simulated task is where a signal + noise input is to be discriminated from a noise-only input. This task is identical to the one used in the experiment of Neri et al. (1998). In the case of biological motion, a number of uncorrelated (but otherwise identical) dots are embedded into a Johansson display. Errors are made in the matching of templates, and a noisy sequence is then processed by the HMM. The response distribution is generated from the log-likelihood output from the HMM. From here it is a relatively simple task to compute a psychometric function and then derive the noise/signal dependence at a fixed performance level.

While many studies in machine vision have made use of HMMs in motion analysis, none has attempted the technique proposed here. These studies have mainly used HMM to model physical motion rather than the decision process itself (e.g., Yamato et al. 1992; Bregler 1997).

3 Design of the model

In this section, we outline the details of the model and show their implementation.

3.1 Motion of structured forms

Figure 4 shows a schematic overview of the proposed computational model for coherent motion detection. The input to the model is a sequence of frames that represent

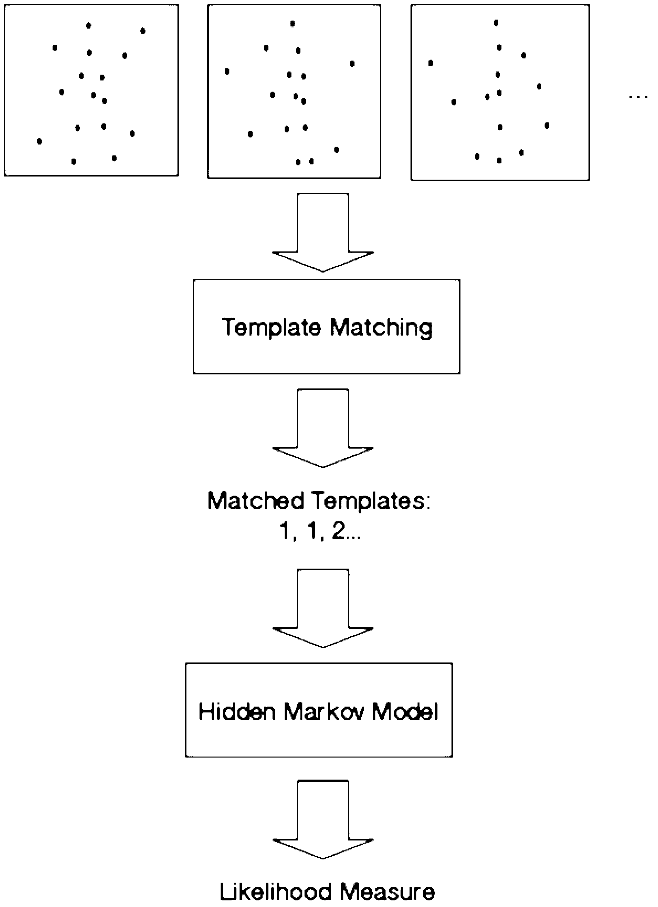

Fig. 4. Schematic diagram illustrating overview of proposed model. Template matching is performed on the input to find the sequence of best match. This sequence is then fed into the HMM, from which a likelihood is calculated. The likelihood is used as a decision variable to decide whether a target is present

a moving light display. The walking person was generated using the algorithm detailed in Cutting (1978). The frame rate of the motion sequence is not a salient parameter, although it is worth mentioning that 36 equally spaced frames were used to digitize one cycle of walking motion.

Four frames were chosen as templates to be used in the model. These choices are shown at the bottom of Fig. 2 and were selected at regular intervals from the walking sequence. The input frames (i.e., a signal + noise frame or noise-only frame) are then individually matched to each of the four template frames (labeled 1–4), and the best match (i.e., highest correlation) was recorded. We employed a distance algorithm that determined the average distance between the dots in the template and the dots in the input frame. Average distance was calculated by a two-dimensional Euclidean metric:

$$D_j = \frac{1}{n_j} \sum_{k \in S_j} \| \mathrm{y}_k - \mathrm{x}_j \| \,, \tag{2}$$

where D_j is the average distance between the jth template point (represented by x_j) and the n_j neighboring points in the input frame (represented by y_k). The set of n_j points lying within a given fixed radial distance from x_j is denoted as S_j. The smaller the average distance, the closer the match. Figure 5 illustrates this algorithm. For each point x_j in the template, points in the input frame within a fixed radial distance of x_j are chosen and the average distance is then calculated. Likewise the average distance is calculated for all template points and the sum total is then used as a measure of correlation. The template with the lowest total distance (i.e., highest correlation) is chosen to best represent the input frame.

The idea of template matching was inspired by the physiological evidence that the neural response in the visual cortex measures the correlation or similarity between the visual stimulus and the optimal pattern the neuron is tuned to recognize (for example, the hand-selective and face-selective cells found in the inferior temporal cortex or IT, Gross et al. 1972; Desimone et al. 1984). Recent studies have suggested that neurons from the so-called "extrastriate body area" can recognize entire bodies (Downing et al. 2001; Grossman and Blake 2002). The use of templates is also supported by the fact that the ventral pathway and higher cortical areas play an important role in the perception of motion of structured forms (Grossman et al. 2000; Vaina et al. 2001).

The sequence of best-matched templates provides the input to the HMM. The HMM was initially trained with all joints present (full signal, no noise) using a standard algorithm known as the forward–backward algorithm (e.g., Poritz 1988; Rabiner 1989). The algorithm optimizes the transition probabilities so that the highest likelihood is assigned to the training pattern. We describe our implementation of the HMM in the appendix.

A predetermined threshold was used to determine the presence or absence of a target. Figure 6 schematically shows the noise-only and signal + noise distributions resulting from a simulation of a forced choice detection task. The noise-only distribution represents the output of the HMM in the case where no target was present (i.e., a noise-only input). A criterion that maximizes the percentage of correct responses was chosen.

3.2 Random coherent motion

To evaluate the model for random coherent motion, we followed the study of Neri et al. (1998) and chose translational motion in a random dot kinematogram. The RDKs were generated with dots that have a limited "lifetime" of one frame. In each frame, the same dots (signal) were displaced horizontally in accordance with translational motion. The remainder of the dots (noise) were placed randomly and are uncorrelated to the signal. Two detection schemes were implemented based on different assumptions regarding the underlying process.

3.2.1 Template method. The first method proposes that high-level processes are involved in the detection of translational motion in RDKs. This method uses the same algorithm discussed earlier for biological motion. However, since the shape of the target is unknown, there can be

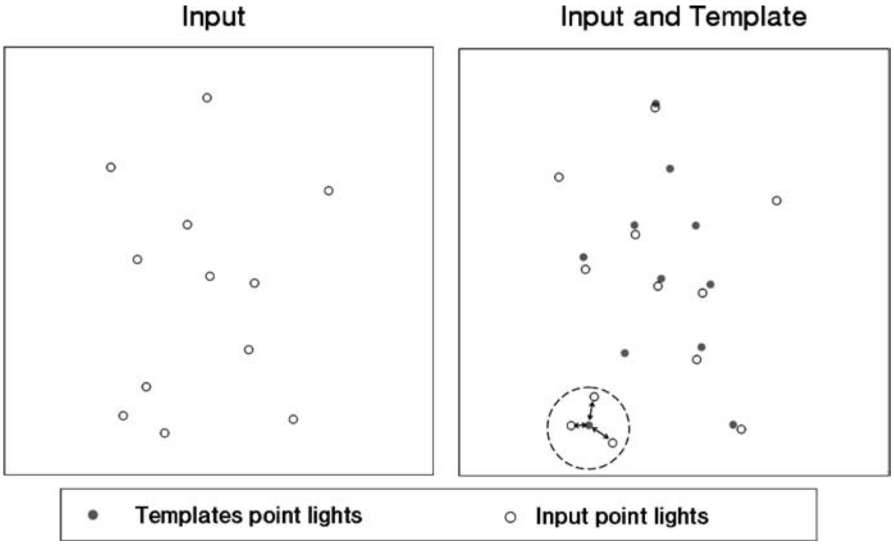

Fig. 5. Illustration of the template-matching algorithm. For each dot in the template, the average distance to all dots within a fixed radial distance is calculated. The average distance is then summed over each dot in the template to obtain a total. This total average distance then defines the distance between the input frame and the template

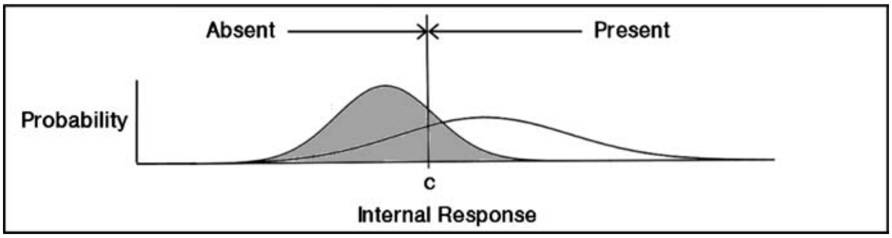

Fig. 6. The signal-detection problem. Two distributions representing the noise-only and signal + noise possibilities are shown. The *vertical line* represents the decision criterion. When the noise-only and signal + noise distributions overlap, the decision task is nontrivial. There are false alarms and false positives in addition to correct responses. Note that this figure is for illustrative purposes only and was not obtained from any simulations

no a priori templates. Any templates used in the detection process must be created "on the fly." Since noise is indistinguishable from signal in the first frame, the entire frame is used as the template. Subsequent templates are then obtained by shifting the first template over by a fixed number of pixels in accordance with the expected pattern of motion. The remainder of the procedure is identical to the method discussed earlier.

3.2.2 Collinear detection method. This method is adapted from a study by Tripathy et al. (1999). The original technique was applied to the detection of static patterns, but we have generalized it to work with translational motion. The method of collinear detection does not involve the use of template matching or HMMs. The likelihood is instead calculated by a different technique. We have provided a more detailed description of the work by Tripathy et al. (1999) in the discussion section. The collinear technique was implemented by taking all input frames and creating a single "superframe" that tracks the succession of dots over time. This "superframe" can be thought of as the entire signal integrated over the duration of the motion, with one important exception: only those dots that advance one position in each frame and follow the sequence expected from translational motion are tracked. A rectangular template is then applied throughout this frame, and the number of times the rectangle is fully occupied is taken to be a measure of the likelihood that the signal is present (Fig. 7).

4 Results

The results of the simulations are discussed. All simulations were carried out with the C programming language and in MATLAB on a PC. We begin with a discussion of the underlying detection distributions.

4.1 The log-likelihood distributions

The signal + noise and noise-only distributions were generated by creating a histogram of the output of the HMM. Consider first the noise-only distribution. Figure 8 shows a typical distribution obtained from a noise-only input. The distribution appears to be Gaussian to a good approximation. This result can be derived analytically in the following manner. When the frames in a noise-only input are uncorrelated, the likelihood from the HMM is calculated from the product of a number of independent probabilities. Taking the logarithm of the likelihood, this product becomes a sum of independent log probabilities. Finally, the central limit theorem guarantees that the distribution of this sum converges asymptotically to a normal distribution (Lee 2003). Figure 8 also illustrates a typical

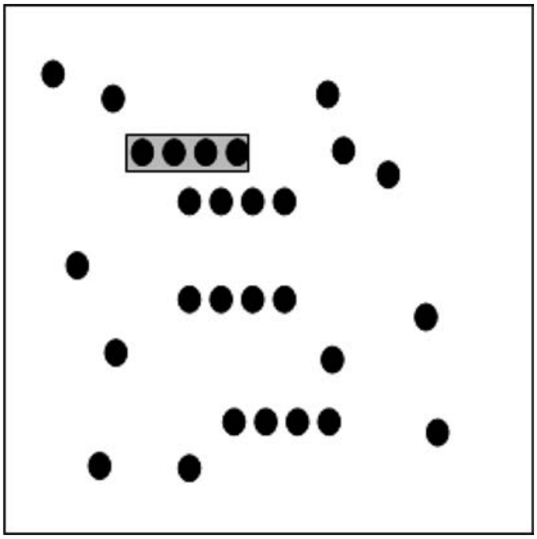

Fig. 7. An example of a "superframe" obtained from the collinear detection method. A rectangular template is used to detect the presence of collinear dots. This example illustrates a case where the total length of the signal is four (i.e., the target dots will move three positions to the right in the stimulus sequence), and hence the size of the rectangular template used to search for a collinear path is also four

distribution for the signal + noise case. The distribution is skewed slightly toward the higher likelihoods. As the amount of noise is increased, the distribution becomes more symmetrical, bearing a closer resemblance to a normal distribution. We have also noted that the variance of the noise-only distribution is generally greater than the signal + noise distribution, giving rise to an asymmetric or unequal variance detection problem.

The distributions for the two translational motion-detection models (template matching and collinear detection) are similar and are not shown here.

4.2 Sensitivity plot

During the initial simulation of the motion-detection task, an adaptive algorithm was used to determine the optimal placement of the threshold. The percentage of correct response was then calculated as a function of noise with the amount of signal held constant. A psychometric function was generated from these results and sufficient data were accrued to obtain unbiased results. The simulations were then repeated for a different value of the signal.

Sensitivity plots were then derived from the psychometric functions. For a fixed performance criterion (say, 80%), the psychometric function defines the average tolerable noise. The noise level was then tabulated as a function of the signal strength. These data were plotted on a double-log plot of signal to noise. The results in all three cases (biological motion + two methods of translational motion) were straight lines, indicating that the underlying dependence between signal and noise follows the equation defined in (1). Please see Fig. 9. The exponent is the most salient parameter here since the intercept is a function of the performance criterion and the parameters of the experiment. For biological motion, the exponent was found to be close to 2 (1.9 obtained from linear regression). For translational motion, both detection methods resulted in an exponent close to 1 (0.88 for template matching and 1.0 for collinear detection).

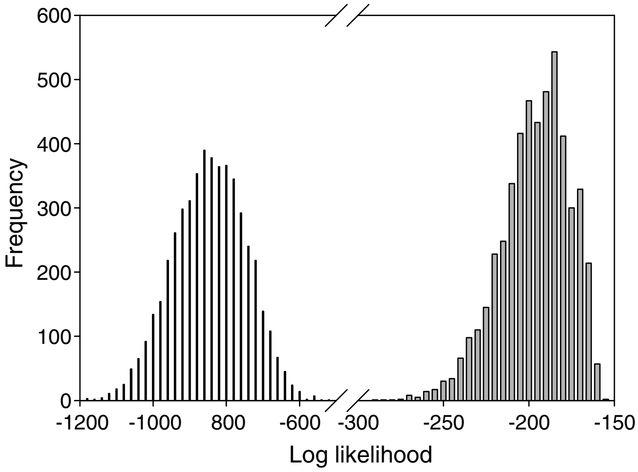

Fig. 8. Typical noise-only (*left*) and signal + noise (*right*) distributions obtained from simulations using a Johansson display. Note that these two distributions were obtained from different runs of the simulation

4.3 Robustness of results

The results presented here were generated from lengthy computer simulations. This limited our ability to explore the model under a wide range of conditions. Nevertheless, in our preliminary study an attempt was made to evaluate the robustness of the results by varying (among other factors) the speed of the walker, number of hidden states used in the HMM, and the performance criterion. In all cases, changes to the value of the exponent in (1) were minimal (<5%). This is a good indication that the values of the exponents are robust.

5 Discussion

This paper proposes a high-level, systems-oriented model of coherent motion detection. The model is based on the idea that the detection of biological or articulated motion is more noise tolerant than motion originating from a source that lacks a definite or recognizable form. At the core of the model is an algorithm for template matching and a hidden Markov process that models the stochastic nature of motion detection. The implications of our model are now discussed.

Central to our approach is the idea that the detection strategy or mechanism will differ according to whether there are stored mental representations or templates of the object to be detected. When the signal is well defined and well known, as in the case of biological motion, templates are used in the detection process. However, when the source of the motion is lacking a definite form or recognizable structure, we cannot rely on existing templates and

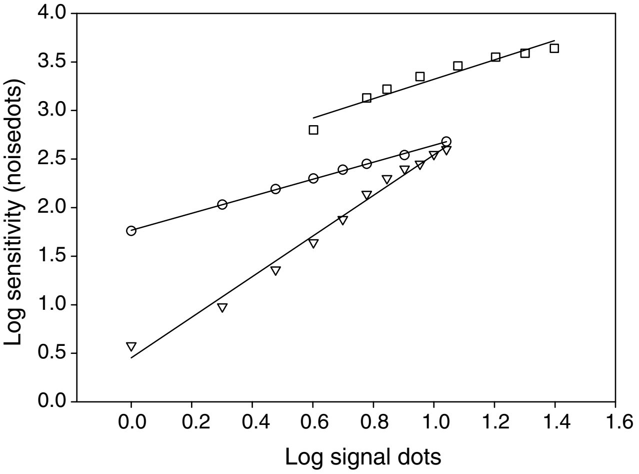

Fig. 9. Sensitivity curves for the Johansson display (*inverted triangles*), translational motion via template matching (*circles*), and translational motion via collinear detection (*squares*). The *solid curves* show the regression lines

must use correspondence mechanisms for the detection. We feel that this is ultimately the reason why the growth in sensitivity is different for biological and random coherent motion.

In the case of biological motion, the use of template matching in our model was inspired by the high-level process of motion perception. The representation of a person walking across a room can be thought of as a collection of images or templates sequenced temporally in a particular order. If the right foot was observed to precede the left foot at an earlier instance, the left foot is expected to overtake the right foot at a later time. Thus the perception of a walking person is based on having the correct sequence of steps match a stored representation. This process is exactly how template matching and HMMs work within the model.

While the focus of this paper has been primarily on spatial factors and their effect on the sensitivity of motion detection, there are a number of other important factors to consider as well. For example, temporal factors like the frame rate and exposure time will also affect the sensitivity of detection. Currently the proposed model does not take these factors into account. For example, we have implicitly assumed that the system is always given sufficient time to process each frame in the input sequence. This is not true in the case of, say, a time-constrained environment where each frame is allocated only a certain amount of processing time. Depending on how the template-matching process is carried out, fewer signal dots are processed, and hence the tolerable noise would be lower. In such a scenario the temporal summation curve (i.e., sensitivity) would rise for increasing exposure time until a plateau is reached where the system is no longer constrained in processing time.

One of the algorithms used in translational motion detection was based upon a method first proposed by Tripathy et al. (1999). In this paper, the authors provided an analysis based upon the so-called "optimal performance condition." Their argument is as follows. Let L and w be, respectively, the length and width of the rectangular template used in the detection of collinear dot patterns in the "superframe" (see earlier description of the collinear detection method). The number of noise dots that fall within the template obeys Poisson statistics with variance LwN, where N is the density of noise dots. The dot density is proportional to the total number of noise dots. In a detection task, the performance measure d' is constant and is defined as

$$d' = \Delta\mu/\sigma = \text{constant}, \tag{3}$$

where $\Delta\mu$ is the difference in mean between the signal + noise and noise-only distributions and σ is the standard deviation. If the detection performance is limited by correspondence noise, σ is obtained from the standard deviation of the Poisson distribution. Hence $\sigma = \sqrt{LwN}$. Since $\Delta\mu$ is equal to the number of target dots S, we have from (3)

$$\frac{S}{\sqrt{LwN}} = \text{constant}. \tag{4}$$

Under optimal performance condition, the length of the template is chosen to be equal to the length of the target. Thus $L = S$. From here a power law emerges with an exponent equal to 1, in agreement with our simulations and with the arguments made earlier in the introduction.

Finally, we note that the model proposed here can be easily extended to more complex systems that allow for the recognition of different types of biological or articulated motion. Figure 5 shows an overview of such a system. The system is comprised of three main components: a hidden Markov model, the template database, and a preprocessor to perform template matching. The template database grows as part of the learning or development process. As templates are required in a particular task, they are loaded into a local cache. Given an arbitrary input to the system, the templates are matched sequentially to the input – the one with the highest likelihood is the pattern that is "recognized" by the system. A system can be trained to recognize a number of different motion types (e.g., walking, running, jumping) by storing a different set of templates for each motion class. Similar maximum likelihood models have been used in speech recognition (e.g., Gold and Morgan 1999) and elsewhere. There are, of course, a number of specific issues that require further consideration before a system can be implemented effectively (e.g., perspective invariance). The solution to this problem might be to include additional preprocessors to help transform the image into an orientation that best matches the stored templates.

6 Conclusions

A high-level, systems-oriented stochastic model of motion detection has been proposed. One consequence of this model is that the detection of biological motion is more noise tolerant than the detection of random coherent motion. Sensitivity curves were generated from simulations, and a power law governing the relation between

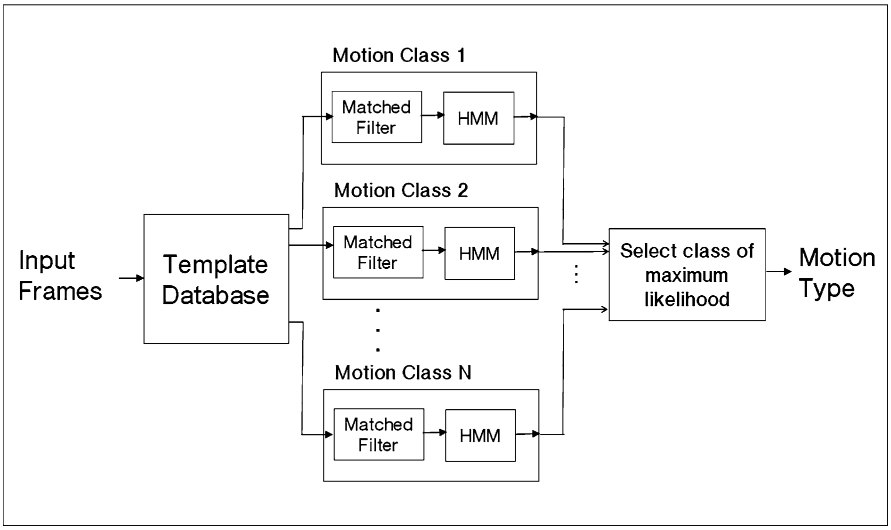

Fig. 10. Outline of a computational vision system that can detect multiple types of articulated, coherent motion. Templates are loaded into a local cache as required. The output of highest likelihood defines the type of motion "seen" by the system

tolerable noise and signal was found. The exponent in the power law was approximately 2 for the detection of biological motion and 1 for the detection of random coherent motion. These preliminary results are in agreement with experimental findings on human subjects, suggesting that the mechanisms of template matching and HMMs may provide a suitable model of the process of motion detection.

Acknowledgements. The authors would like to thank Professor Horace Barlow for his many helpful comments and suggestions. This work was supported by the National Science and Engineering Research Council of Canada (NSERC).

Appendix: Details concerning the implementation of the hidden Markov models

The HMMs used in this study were implemented using the Hidden Markov Model Toolbox for MATLAB (Murphy 2003). Each HMM had four hidden and three observable states. An HMM is defined by the initial state probabilities, the transition probabilities, and the observation probability that map the hidden states onto the observable states. The HMMs were trained using the standard forward–backward algorithm. This algorithm is carried out iteratively and is comprised of the following steps: (1) randomize all parameters; (2) calculate forward and backward probabilities according to:

$$\alpha_t(s)=\sum_{r\in S}\alpha_{t-1}(r)a_{rs}b_s(O_t)\,, \tag{5}$$

$$\beta_t(s)=\sum_{r\in S}a_{sr}b_r(O_{t+1})\beta_{t+1}(r)\,, \tag{6}$$

where a_{sr} is the probability that governs the transition from state s to r, $b_s(O_t)$ is the probability of the observation at time t given hidden state s, and S is the set of all possible hidden states; (3) calculate posterior probabilities and state transitions from

$$\gamma_t(s)=\alpha_t(s)\beta_t(s)\Big/\sum_{s\in S}\alpha_T(s)\,, \tag{7}$$

$$\gamma_t(r,s)=\alpha_t(r)a_{r,s}b_s(O_{t+1})\beta_{t+1}(s)\Big/\sum_{s\in S}\alpha_T(s)\,, \tag{8}$$

where $\gamma_t(s)$ is the posterior probability of state s at time t, $\gamma_t(r,s)$ the posterior probability of transition from state r to s at time t, and T the total number of frames; (4) calculate initial, transition, and observation probabilities:

$$\pi_s=\gamma_1(s) \tag{9}$$

$$a_{r,s}=\sum_{t=1}^{T-1}\gamma_t(r,s)\Big/\sum_{s'\in S}\sum_{t=1}^{T-1}\gamma_t(r,s') \tag{10}$$

$$b_s(k)=\sum_{t:O_t=k}\gamma_t(s)\Big/\sum_{t=1}^{T}\gamma_t(s); \tag{11}$$

and (5) repeat steps 2, 3, and 4 until convergence is satisfied. The initial training sequence was obtained with template matching on a noise-free walking sequence.

References

Adelson EA, Bergen JR (1985) Spatio-temporal energy models for the preception of motion. J Opt Soc Am A 2:284–299

Barlow HB, Tripathy SP (1997) Correspondence noise and signal pooling in the detection of coherent motion. J Neurosci 17:7954–7966

Braddick OB, O'Brien JM, Wattam-Bell J, Atkinson J, Hartley T, Turner R (2000) Brain areas sensitive to coherent visual motion. Perception 30:61–72

Bregler C (1997) Learning and recognizing human dynamics in video sequences. In: Proceedings of IEEE conference on computer vision and pattern recognition, San Juan, Puerto Rico, 17–19 June 1981, pp 568–574

Britten KH, Shadlen MN, Newsome WT, Movshon JA (1992) The analysis of visual motion: a comparison of neuronal and psychophysical performance. J Neurosci 12:4745–4765

Brownlow S, Dixon AR, Egbert CA, Radcliffe RD (1997) Perception of movement and dancer characteristics from point-light displays of dance. Psychol Rec 47:411–421

Celebrini S, Newsome WT (1994) Neuronal and psychophysical sensitivity to motion signals in extrastriate area MST of the Macaque monkey. J Neurosci 14:4109–4124

Cutting JE (1978) A program to generate synthetic walkers as dynamic point-light displays. Behav Res Meth Instrum 10:91–94

Cutting JE, Proffitt DR (1981) Gait perception as an example of how we may perceive events. In: Walk R, Pick HL (eds) Intersensory perception and sensory integration. Plenum, New York, pp 249–273

Dawson MR (1991) The how and why of what went where in apparent motion: modeling solutions to the motion correspondence problem. Psychol Rev 98:569–603

Desimone R, Albright TD, Gross CG, Bruce C (1984) Stimulus-selective properties of inferior temporal neurons in the macaque. J Neurosci 4:2051–2062

Dittrich WH (1993) Action categories and the perception of biological motion. Perception 22:15–22

Downing PE, Jiang Y, Shuman M, Kanwisher N (2001) A cortical area selective for visual processing of the human body. Science 293:2470–2473

Fredericksen RE, Verstraten FA, van de Grind WA (1993) Spatio-temporal characteristics of human motion perception. Vis Res 33:1193–1205

Giese MA, Poggio T (2003) Nerual mechanisms for the recognition of biological movements. Nat Rev Neurosci 4:179–192

Goddard N (1992) The perception of articulated motion: recognizing moving light displays. PhD Thesis, University of Rochester, Rochester, NY

Gold B, Morgan N (1999) Speech and audio signal processing: processing and perception of speech and music. Wiley, New York

Green DM, Swets JA (1966) Signal detection theory and psychophysics. Krieger, New York

Gross CG, Rocha-Miranda CE, Bender DB (1972) Visual properties of neurons in inferotemporal cortex of the macque. J Neurophysiol 35:96–111

Grossman ED, Blake R (2002) Brain areas active during visual perception of biological motion. Neuron 35:1167–1175

Grossman ED, Donnelly M, Prices R, Pickens D, Morgan V, Neighbor G, Blake R (2000) Brain areas involved in perception of biological motion. J Cogn Neurosci 12:711–720

Grzywacz NM, Watamaniuk SNJ, McKee SP (1995) Temporal coherence theory for the detection and measurement of visual motion. Vis Res 35:3183–3203

Johansson G (1973) Visual perception of biological motion and a model for its analysis. Percept Psychophys 14:201–211

Kozlowski LT, Cutting JE (1977) Recognizing the sex of a walker from a dynamic point-light display. Percept Psychophys 21:575–580

Lee J (2003) A computational model for biological motion perception. Master's Thesis, University of Toronto

Marr D (1982) Vision. Freeman, San Francisco

Mather G, Murdoch L (1994) Gender discrimination in biological motion displays based on dynamic cues. Proc R Soc Lond B 258:273–279

Murphy K (2003) Hidden Markov Model (HMM) Toolbox. http://www.ai.mit.edu/ murphyk/Software/HMM/hmm.html

Neri P, Morrone MC, Burr DC (1998) Seeing biological motion. Nature 395:894–896

Newsome WT, Britten KH, Movshon JA (1989) Neuronal correlates of a perceptual decision. Nature 341:52–54

Peterson MA, Gibson BS (1991) The initial identification of figure-ground relationships: contributions from shape recognition processes. Bull Psychon Soc 29:199–202

Poritz AB (1988) Hidden Markov models: a guided tour. In: Proceedings of the IEEE conference on acoustics, speech and signal processing1. IEEE Press, New York, 1:7–13

Rabiner LR (1989) A tutorial on hidden Markov models and selected applications in speech recognition. Proc IEEE 77:257–286

Tripathy SP, Mussap AJ, Barlow HB (1999) Detecting collinear dots in noise. Vis Res 39:4161–4171

Troje NF (2002) Decomposing biological motion: A framework for analysis and synthesis of human gait patterns. J Vis 2:371–387

Vaina LM, Solomon J, Chowdhury S, Sinha P, Belliveau JW (2001) Functional neuroanatomy of biological motion perception in humans. Proc Nat Acad Sci USA 98:11656–11661

Watamaniuk SN, McKee SP, Grywacz NM (1995) Detecting a trajectory embedded in random-direction motion noise. Vis Res 35:65–77

Wong W, Barlow HB (2000) Tunes and templates. Nature 404:952–953

Yamato J, Ohya J, Ishii K (1992) Recognizing human action in time-sequential images using hidden Markov model. In: Proceedings of IEEE conference on computer vision and pattern recognition, Champaign, IL, 15–18 June 1992, pp 379–385