A Computational Model for Motion Detection and Direction Discrimination in Humans

Yang Song† and Pietro Perona†‡
† California Institute of Technology, 136-93, Pasadena, CA 91125, USA
‡ Università di Padova, Italy
{yangs,perona}@vision.caltech.edu

Abstract

Seeing biological motion is very important for both humans and computers. Psychophysics experiments show that the ability of our visual system for biological motion detection and direction discrimination is different from that for simple translation [4]. But the existing quantitative models of motion perception can not explain these findings. We propose a computational model, which uses learning and statistical inference based on the joint probability density function (PDF) of the position and motion of the body, on stimuli similar to [4]. Our results are consistent with the psychophysics indicating that our model is consistent with human motion perception, accounting for both biological motion and pure translation.

1. Introduction

Perceiving the motion of the human body ('biological motion' in the literature of human vision) is a most important ability for the human visual system. Understanding how the brain perceives human motion and developing a computational model for it is an interesting and challenging problem for the fields of computer vision and human vision.

The abilities of the human visual systems for detection and direction discrimination, for both simple translation and biological motion, have been measured psychophysically [4]. In [4], a Johansson-like display [3] was used and it is found that the ability of the visual system to integrate biological motion over space and time is different from that of simple translation. Sensitivity to biological motion increases rapidly with the number of displayed joints, far more rapidly than for translation.

Many quantitative models of motion perception have been proposed, for example those in [5, 1, 9]. But they have been developed for translation, not for biological motion. No existing computational model can explain the difference between biological motion perception and translation perception as found in [4].

In [7, 6, 8], we proposed a perceptual model for detecting a moving human and for labeling its parts automatically. Rather than modeling the details of the mechanics of the human body, we choose to approach biological motion perception as the problem of recognizing a peculiar spatio-temporal pattern which may be learned perceptually. We observe the subject moving about in order to estimate a model of his/her stereotypical motions. This model, which we formulate as the joint probability density function (PDF) of the position and motion of the body, has a Markov-like structure.

The above model has demonstrated excellent and efficient performance on motion sequences with clutter and occlusion. It is therefore very interesting to compare the performance of it with that of the human visual system and examine if it can model how the human visual system behaves. In this paper we apply the probabilistic model to the tasks of detection and direction discrimination using stimuli similar to [4] and compare the results with the psychophysics results.

In section 2, the tasks and stimuli used to test the model are depicted. The probabilistic model is explained in section 3. Section 4 contains our simulation results, which are compared with psychophysics experiments in section 5.

2. Our stimuli

There are two kinds of tasks: one is to detect the presence of the target, and the other is to discriminate the direction of the target motion, both in the presence of dynamic random noise. The target is either a walker (biological motion) or simple translation. In the following, we first describe how the signals (targets) are generated, and then explain the tasks in more details.

0-7695-0939-8/00 $10.00 © 2000 IEEE

2.1. Signals

Biological motion: We use the same program as in [4, 2] to generate the human walking sequence, where the motion of 13 dots represents the motion of the main joints of a person walking on a treadmill. Since we want to study how performance changes with the number of displayed joints, only a subset of the 13 joints appear in each frame. Each signal dot has a 'limited-lifetime' of two frames, then is 'reborn' at a randomly chosen joint, that is, we randomly select which joints to be displayed for each pair of frames, and during the whole sequences each joint has an equal chance to be represented.

Translation: Signal dots are generated in random positions over the area of the walker, with all moving at the same speed (set to match the average speed of the individual dots of biological motion). As in biological motion, the lifetime of each dot is also assumed to be two. Therefore, the positions of signal dots are generated randomly for the first frame, then those dots move to the second frame, and the positions are generated randomly again for the third frame, and so on.

2.2. Tasks

Detection. Detection is to decide which one of the two side by side displays contains the target: one consists of signal dots with certain amount of noise dots, and the other is a control display with the same dots density. Noise dots are generated independently for each frame using a uniform probability density. For biological motion, the control dots are derived from the walking algorithm [2] by randomizing the order of the frames presented. For translation, the control dots are generated independently for each frame.

Direction discrimination. Direction discrimination is to determine whether the target is moving rightwards or leftwards for a display known to contain the target. The display consists of signal dots superimposed with dynamic noise dots. The signal dots are generated as in section 2.1 for either biological motion or translation, and noise dots are generated randomly for each frame.

3. Computational Model

In the following subsections, the approaches of doing detection and direction discrimination from two frames are described. Based on the results from two frames, decisions upon multiple frames can be made handily as in [8].

For a pair of frames, positions and velocities of point features are taken as measurements, which are obtained from the local maxima of the Reichardt motion energy [5, 1, 9] between the two frames (see Appendix for our implementation).

3.1. Detection

Given two sets of measurements $\overline{X}_1$ and $\overline{X}_2$, detection is to decide which of the following two hypotheses is true:

Hypothesis 1 (O_1): $\overline{X}_1$ contains the target;

Hypothesis 2 (O_2): $\overline{X}_2$ contains the target.

Therefore, if $P(O_i|\overline{X}_i)$, $i = 1, 2$, is the posterior probability of the hypothesis O_i given $\overline{X}_i$, we need to compute the ratio

$$\begin{aligned} R(\overline{X}_1, \overline{X}_2) &= \frac{P(O_1|\overline{X}_1)}{P(O_2|\overline{X}_2)} \\ &= \frac{P(\overline{X}_1|O_1)P(O_1)/P(\overline{X}_1)}{P(\overline{X}_2|O_2)P(O_2)/P(\overline{X}_2)} \\ &= \frac{P(\overline{X}_1|O_1)}{P(\overline{X}_2|O_2)} \cdot \frac{P(O_1)}{P(O_2)} \cdot \frac{P(\overline{X}_2)}{P(\overline{X}_1)} \end{aligned} \quad (1)$$

where the second equal sign holds according to Bayes' law. If $R(\overline{X}_1, \overline{X}_2)$ is greater than 1, then $\overline{X}_1$ contains the target; otherwise the target is in $\overline{X}_2$. If the prior probabilities are assumed to be equal, the last two terms of the above equation are 1. As in [7, 8, 6], let $\overline{L}$ denote a possible labeling of $\overline{X}_1$ and assume $\mathcal{L}$ is all the possible labelings when $\overline{X}_1$ contains the target (O_1), then

$$\begin{aligned} P(\overline{X}_1|O_1) &= \sum_{\overline{L}\in\mathcal{L}} P(\overline{X}_1, \overline{L}|O_1) \\ &= \sum_{\overline{L}\in\mathcal{L}} P(\overline{X}_1|\overline{L}, O_1)P(\overline{L}|O_1) \end{aligned} \quad (2)$$

If we don't have any prior information about the labeling, then we can assume in the above equation, for any labeling $\overline{L}$, $P(\overline{L}|O_1) = 1/|\mathcal{L}|$, where $|\mathcal{L}|$ is the number of possible labelings. Let $\overline{X}_{fg}$ denote the foreground (target) measurements in $\overline{X}$, $\overline{X}_{bg}$ the measurements of background features, and $\overline{X}_{fg} \cup \overline{X}_{bg} = \overline{X}$. If foreground measurements and background measurements are independent,

$$P(\overline{X}_1|\overline{L}, O_1) = P_{fg}(\overline{X}_{fg}) \cdot P_{bg}(\overline{X}_{bg}) \quad (3)$$

If independent uniform background noise is assumed, $P_{bg}(\overline{X}_{bg})$ can be computed easily [8]. The computation of $P_{fg}(\overline{X}_{fg})$ will be described in section 3.3. $P(\overline{X}_2|O_2)$ can similarly be obtained.

3.2. Direction discrimination

For a given set of measurements $\overline{X}$, direction discrimination is to decide between the following two hypotheses:

Hypothesis 1 (H_1): rightward motion

Hypothesis 2 (H_2): leftward motion

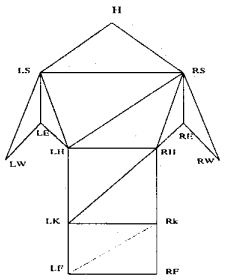

Figure 1. Decomposition of the human body into triangles [7]. 'L' and 'R' in label names indicate left and right. H:head, N:neck, S:shoulder, E:elbow, W:wrist, H:hip, K:knee and F:foot.

If $P(H_1|\overline{X}) > P(H_2|\overline{X})$, then the motion is rightwards, and vice versa.

$$\frac{P(H_1|\overline{X})}{P(H_2|\overline{X})} = \frac{P(\overline{X}|H_1)P(H_1)/P(\overline{X})}{P(\overline{X}|H_2)P(H_2)/P(\overline{X})} = \frac{P(\overline{X}|H_1)}{P(\overline{X}|H_2)} \cdot \frac{P(H_1)}{P(H_2)} \quad (4)$$

If we assume the prior probabilities are equal, i.e. $P(H_1) = P(H_2)$, then the decision rule becomes if $P(\overline{X}|H_1) > P(\overline{X}|H_2)$, the motion is rightwards, and vice versa. $P(\overline{X}|H_1)$ and $P(\overline{X}|H_2)$ can be computed in a similar way to equations (2) and (3).

3.3. Triangulated model for foreground probability

Biological motion [7, 8]. We first consider the case where all the body parts are present. By using the kinematic chain structure of human body, the whole body can be decomposed as in Figure 1. If the appropriate conditional independence (Markov property) is valid, then

$$P_{fg}(\overline{X}_{fg}) = P_{LW,LE,LS}(X_{LW}|X_{LE},X_{LS})P_{LE,LS,LH}(X_{LE}|\dots) \dots P_{RK,LF,RF}(X_{RK},X_{LF},X_{RF}) \quad (5)$$

where LW is the left wrist, RF is the right foot, etc; X_{LW} is the measurements (positions and velocities) of left wrist, X_{RF} is the measurements of right foot, etc. For our stimuli with some body parts missing in each frame, the foreground probability $P_{fg}(\overline{X}_{fg})$ is the marginalized version of equation (5) and can be computed as in [8]. Under this triangulated decomposition, the summation in equation (2) can be computed in polynomial time (on the order of N^4 where N is the number of observed features).

Translation. In case of no features missing, by the way of translation stimuli generated, the total number of signal dots in a translation display is the same as that of biological motion (13 here). To test the model, we assume that the the joint foreground probability density function (PDF) for translation can also be decomposed into multiplications of joint (or conditional) PDF of triplets as in equation (5) , i.e.,

$$P_{fg}(\overline{X}_{fg}) = P_{A,B,C}(X_A|X_B,X_C)P_{B,C,D}(X_B|X_C,X_D) \dots P_{K,L,M}(X_K,X_L,X_M) \quad (6)$$

where $A, B, \dots, L, M$ are 13 labels, and $X_A, X_B, \dots, X_M$ are the corresponding measurements. Similarly to biological motion, when some features are missing (the number of signal dots is less than 13), the marginalized version ([8]) of the equation (6) is used to compute $P_{fg}(\overline{X}_{fg})$.

Though the probabilistic model structure and the computing method are the same for biological motion and translation, the model parameters (e.g. mean and covariance for Gaussian PDF) are different because the training sets are different (as will be explained in the next section).

4. Experiments

In our experiments, the probabilistic models are first learned, and then applied to the stimuli as described in section 2.

4.1. Training of the probabilistic models

Four kinds of foreground probabilistic model are learned: rightward and leftward motion for biological motion and translation respectively.

Biological motion. The training sequence is generated by the program in [4, 2]. For each pair of frames, positions and velocities are taken as measurements. Since the ground truth (labeled data) is needed for training, the velocities are obtained by subtracting the positions in two consecutive frames, not from Reichardt model. Independent uniform noise is added to both positions and velocities to match the quantization error introduced by the Reichardt detector which is used calculating velocities in the test data.

The training was done by estimating the joint (or conditional) probabilistic density functions (pdf) for all the triplets as described in section 3.3. As in [8, 6], we assumed all the pdfs were Gaussian, and the parameters for the Gaussian distribution were estimated from the training set.

Translation. For each frame, the positions of signal dots are generated randomly over the area, and the velocities are assigned to be the same for the dots in the same frame and generated from a uniform distribution over a certain range (identical to the range observed in biological motion, horizontal only and opposite for rightward and leftward model) across frames. Independent uniform noise is added to velocities to simulate the quantization error from Reichardt model.

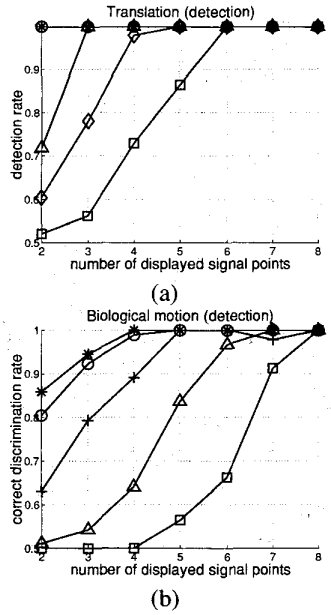

Figure 2. **Detection rate vs. number of displayed signal points (joints) under several noise levels.** (a) translation; (b) biological motion. The noise levels are: square 200 noise dots; diamond 150 noise dots (for translation only); triangle 100; plus 50; circle 30; star 10.

To use the decomposition model as in equation (6) for translation, labels are assigned randomly to signal dots for each pair of frames. The joint (or conditional) PDFs for all the triplets are assumed to be Gaussian.

4.2. Detection

The detection task, for both biological motion and translation, is performed on stimuli as in section 2. The size of the display is 170 by 310 pixels, a set-up very close to that in [4].

The algorithm described in section 3.1 is used. $\overline{X}_1$ and $\overline{X}_2$, which are positions and velocities for the image containing target and the control image, are obtained through Reichardt energy model so that $P(O_1|\overline{X}_1)$ and $P(O_2|\overline{X}_2)$ can be computed. In our simulations, we integrate over 5 pairs of frames to make decisions [8].

To compare with psychophysics results in [4], we study how sensitivity varies with the number of displayed signal dots (joints). As in [4], sensitivity is defined as the noise level (number of noise dots) at which 75% correct decisions are made. To find the sensitivity for a certain number of displayed signal dots, several noise levels were tried, and sensitivity was calculated by fitting a raised cumulative Gaussian curve (with asymptotes at 0.5 and 1) to the psychometric functions.

Figure 2 shows the detection rate vs. the number of displayed signal points (joints) under several noise levels, for (a) translation and (b) biological motion. For each condition (with a certain number of displayed signal points and a certain noise level), 360 frames (3 gait cycles) were used. The star (with dashed line) in Figure 3, derived from Figure 2(a), is the log-log sensitivity vs. number of displayed signal dots curve for translation detection, with a slope of 0.95 (calculated by linearly line fitting). The square (with solid line) in Figure 3, obtained from Figure 2(b), is the log-log curve for biological motion detection, with a slope of 1.53.

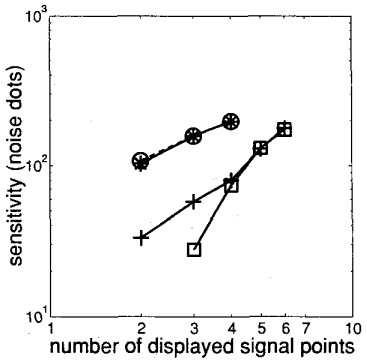

Figure 3. **Sensitivity vs. number of displayed joints.** Star (with dashed line): translation detection; circle (with dashed line): translation direction discrimination; plus (with solid line): biological motion detection; square (with solid line): biological motion direction discrimination.

4.3. Direction discrimination

The direction discrimination task assumes that a moving target is in the scene and needs to decide the direction of the motion. For a given pair of images, positions and velocities are obtained using Reichardt model as measurements ($\overline{X}$), and plugged into $P(\overline{X}|H_1)$ and $P(\overline{X}|H_2)$ as described in section 3.2. As in detection, decisions are then made upon integration over 5 pairs of frames. Sensitivities are calculated in the same way as in section 4.2.

The curves in Figure 4 are the correct direction discrimination rate vs. the number of displayed signal points (joints) under several noise levels, for (a) translation and (b) biological motion. The circle (with dashed line) in Figure 3, derived from Figure 4(a), is the log-log curve of sensitivity vs. number of displayed signal dots for translation direction discrimination, with a slope of 0.88. The plus (with solid line) in Figure 3, obtained from Figure 4(b), is the log-log curve for biological direction discrimination, with a slope of 2.71.

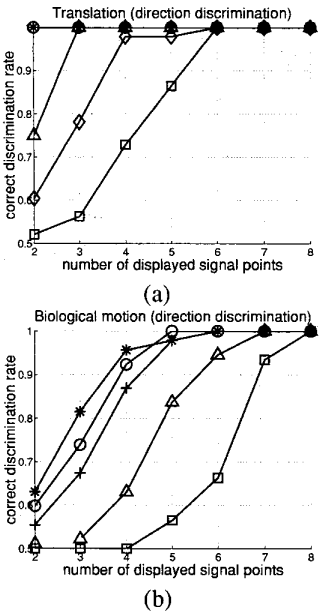

Figure 4. **Correct direction discrimination rate vs. number of displayed signal points (joints) under several noise levels. (a)** translation; **(b)** biological motion. The noise levels are: square 200 noise dots; diamond 150 noise dots (for translation only); triangle 100; plus 50; circle 30; star 10.

5. Comparison with psychophysics

From Figure 3, the log-log curves for perceiving biological motion are steeper than those for translation, that is, the performance of biological motion perception changes more rapidly with the number of displayed signal points than that of translation, which is consistent with the psychophysics results in [4]. This indicates that the number of displayed signal points is more crucial to biological motion perception, which implies that to perceive biological motion, more concerted movement (collaboration) among signal dots is needed. The steepest curve is the one of biological motion direction discrimination, which conforms with our intuition (also the results of psychophysics) that the direction perception of human motion is the most demanding task and needs the signal dots to be the most concerted (collaborative).

The above observation can be explained intuitively by our probabilistic model. For biological motion, the relative positions and velocities are correlated in the PDF of one triplet. But for translation, relative positions and velocities are independent, velocities of different parts are highly correlated and with small variance, and positions are almost independent and with large variance. Therefore, for biological motion, if only two signal points with big relative distance (not in the same triangle) are observed, it is very unlikely for our probabilistic model to take them as a human configuration. For translation, if two signal points are observed, then regardless of their relative position, they can give a higher likelihood being translation as long as they have similar velocities. So in some sense, when the number of signal dots is small, the dots of translation are more 'informative' than those of biological motion.

Our experimental set-up is very similar to that of [4], but different from theirs in the temporal integration part. In their paper, they used the 'limited-lifetime' technique and integration over 1200ms (40 frames). In our temporal integration, we assume independence among pairs of frames, and only integrate over 5 pairs. We believe that experiments with the same condition as them would be qualitatively similar.

6. Conclusions

The consistency between our results and the psychophysics both of biological motion and translational motion perception suggests that our model could be a good computational model for human motion perception.

Our probabilistic model indicates that the visual system may gain the ability of perceiving biological motion and translation through learning. The mechanisms for perceiving biological motion and translation could be the same, but are tuned to different model parameters. When biological motion is perceived, it may not be viewed as a whole, but some closer (or more correlated) body parts may be grouped together first.

Our model could predict the performance of the human visual system on any complex motion pattern. Detailed comparison of such predictions with the psychophysics would allow further refinements of the model.

Appendix: implementation of Reichardt-type feature velocities between two frames [5, 1, 9]

This appendix describes our implementation of getting point feature velocities between two frames using a Reichardt-type model. A image sequence can be represented as a function $I(x, y, t)$, where x and y are spatial coordinates in horizontal and vertical directions respectively, and t is the time coordinate. We compute the velocities between two frames $I(x, y, t)$ and $I(x, y, t+1)$ (for simplicity the time interval is assumed to be 1) in three steps:

(1) Spatial filtering is first applied to both images. Let K(x,y) denote the filter (we use the same filter for both images), then,

$$f_1(x, y) \stackrel{\text{def}}{=} I(x, y, t) * K(x, y)$$
$$f_2(x, y) \stackrel{\text{def}}{=} I(x, y, t+1) * K(x, y)$$

where $*$ means convolution, and $f_1(x, y)$ and $f_2(x, y)$ are the two images after spatial filtering.

(2) Get motion energy under different velocities. Let vx and vy be respectively the horizontal and vertical velocities between the two frames, then $E(x, y, vx, vy)$, which is the motion energy for velocity (vx, vy) at location (x, y), is computed as

$$E(x, y, vx, vy) = f_1(x, y) \cdot f_2(x + vx, y + vy)$$

(3) The local maxima of $E(x, y, vx, vy)$ are taken as the feature velocities between the two frames, that is, velocity (vx_i, vy_i) can be perceived at location (x_i, y_i) if

$$\frac{\partial E}{\partial x}|_{(x_i, y_i, vx_i, vy_i)} = 0, \qquad \frac{\partial E}{\partial y}|_{(x_i, y_i, vx_i, vy_i)} = 0,$$
$$\frac{\partial E}{\partial vx}|_{(x_i, y_i, vx_i, vy_i)} = 0, \quad \text{and} \quad \frac{\partial E}{\partial vy}|_{(x_i, y_i, vx_i, vy_i)} = 0$$

(x_i, y_i, vx_i, vy_i)'s are positions and velocities of point features between the two frames.

Note that in our implementation, x, y, vx and vy are all discretized, therefore, the resolution of the features (x_i, y_i, vx_i, vy_i) depends on the quantization scale. Also, energy E is only computed for a certain range of (vx, vy), which limits the range of a feature velocity (vx_i, vy_i) can be in.

Acknowledgments

Funded by the NSF Engineering Research Center for Neuromorphic Systems Engineering (CNSE) at Caltech (NSF9402726), and by an NSF National Young Investigator Award to PP (NSF9457618). We thank Peter Neri for providing the code of generating the human walking sequences [2].

References

[1] E. Adelson and J. Bergen. Spatiotemporal energy models for the perception of motion. *J. Opt. Soc. Am. A*, 2:284–299, 1985.

[2] J. Cutting. A program to generate synthetic walkers as dynamic point-light displays. *Behav. Res. Methods Instrument*, 10:91–94, 1978.

[3] G. Johansson. Visual perception of biological motion and a model for its analysis. *Perception and Psychophysics*, 14:201–211, 1973.

[4] P. Neri, M. Morrone, and D. Burr. Seeing biological motion. *Nature*, 395:894–896, 1998.

[5] W. Reichardt. Autocorrelation, a principle for the evaluation of sensory information by the central nervous system. In *Sensory Communication, W.A. Rosenblith, ed.* Wiley, New York, 1961.

[6] Y. Song, X. Feng, and P. Perona. Towards detection of human motion. In *Proc. IEEE CVPR*, volume 1, pages 810–817, June 2000.

[7] Y. Song, L. Goncalves, E. D. Bernardo, and P. Perona. Monocular perception of biological motion - dection and labeling. In *International Conference on Computer Vision*, pages 805–812, Sept 1999.

[8] Y. Song, L. Goncalves, and P. Perona. Monocular perception of biological motion - clutter and partial occlusion. In *Proc. ECCV*, volume 2, pages 719–733, June/July 2000.

[9] J. van Santen and G. Sperling. Elaborated reichardt detectors. *J. Opt. Soc. Am. A*, 2:300–321, 1985.