

Visual perception of biological motion by form: A template-matching analysis

Joachim Lange

Department of Psychology, Institute II, Westf. Wilhelms-University, Münster, Germany



Karsten Georg

Department of Psychology, Institute II, Westf. Wilhelms-University, Münster, Germany



Markus Lappe

Department of Psychology, Institute II, Westf. Wilhelms-University, Münster, Germany



Biological motion perception is referred to as the ability to recognize a moving human figure from no more than a few moving point lights. Such point-light stimuli contain limited form information about the shape of the body and local image motion signals from the moving points. The contributions of form and motion to the vivid perception of point-light displays are subject to controversy in the discussion. While some studies claim that local motion signals are critical, others emphasize the role of global form cues. Here, we present a template-matching approach to investigate the role of global form analysis. We used a template-matching method that derives biological motion exclusively from form information. The algorithm used static postures monitored from walking humans as stored templates. We compared the simulation results to psychophysical experiments with the commonly used point-light walker and a variant point-light walker with near-absent local motion signals. The common result in all experiments was a high correlation between simulation results and psychophysical data. The results show that the limited form information in point-light stimuli might be sufficient to perceive biological motion. We suggest that it is possible for humans to extract the sparse form information in point-light walkers and to use it to perceive biological motion by integrating dynamic form information over time.

Keywords: biological motion, form recognition, motion, template-matching analysis

Introduction

Perceiving human movements is a complex task for the visual system because human movements contain many degrees of freedom and involve both rigid and nonrigid elements. Yet, naive human observers readily recognize moving human figures and their complex actions within fractions of a second. This is true even if the stimulus is degraded to only 12 point lights attached to the joints on the body (Johansson, 1973). This striking phenomenon is referred to as “perception of biological motion.”

Biological motion contains different kinds of motion and form information (Figure 1). Each light point changes position over time and thus provides apparent motion signals. We call these the “local” or “image” motion signals. The instantaneous positions of all light points at any time provide structural information about the momentary posture of the body. Although this information is only weak in a single snapshot of a human body, temporal integration of the instantaneous position signals over a sequence of postures may provide increased structural information. We call this the “global form” information. Changes of the structural information of the body posture over time also provide motion information. In this article,

this is referred to as “global motion” information (Figure 1).

The perceptual origin of global motion impressions is still an issue of discussion. Beintema and Lappe (2002) investigated whether normal observers can perceive biological motion in the absence of image motion. They developed a stimulus, which consisted of a fixed number of dots spread randomly over the skeleton of a human figure. The dots were reallocated to a new position every n th frame. For $n = 1$, the position was changed for each frame, thus minimizing useful local image motion information in the stimulus. By varying n , the contribution of local image motion signals could be manipulated (see Stimuli section for details). Spontaneous recognition of this new stimulus by naive observers was similar to that of the classical Johansson stimulus. In various discrimination experiments, Beintema and Lappe and Beintema, Georg, and Lappe (in press) investigated more precisely the role of form information and image motion signals. They manipulated the amount of form information by changing the number of simultaneously visible dots. The results revealed a clear relationship between available form information and discrimination performance of the subjects. Adding local motion signals, on the other hand, did not improve the subjects’ performance, and, in fact, their performance deteriorated marginally. Beintema and Lappe

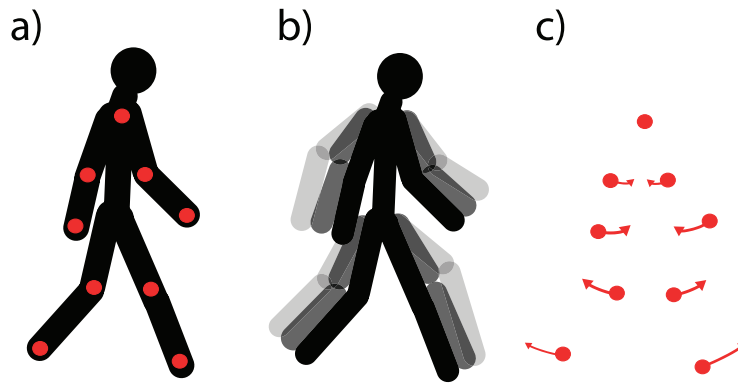


Figure 1. (a) The shape of a human figure contains global (illustrated by the black sketch) and local (illustrated by the red dots) features. (b) The impression of a walking person may occur from the integration of the global shape over time (differently shaded figures) or (c) by integrating the local image motion signals (illustrated by arrows).

suggested that biological motion perception might be achieved by an analysis of the dynamic form of the human figure and that image motion signals have a supporting character in more complicated tasks and are not essential for perception.

The importance of form cues for biological motion perception has also been demonstrated in earlier studies. Chatterjee, Freyd, and Shiffrar (1996) studied the perception of apparent motion from sequential full-body images and found a higher level of usage for biomechanically consistent motion paths compared with impossible motion paths. This motion percept relates to the global motion of the body and overwrites local apparent motion signals when there is a conflict between the two. In another study, Shiffrar, Lichtey, and Heptula Chatterjee (1997) report an orientation-specific recognition of biological motion through apertures, although other objects could not be identified in this manner. Both studies support a role of global form mechanisms for biological motion perception. Because they used line drawings or full-body photographs, the question whether global form analysis can also explain biological motion perception from point-light stimuli remains open.

Bertenthal and Pinto (1994) investigated the importance of form for the recognition of point-light biological motion. Using masks comprising dots with trajectories identical to those of the walker itself but with different, randomly chosen positions, they concluded that biological motion perception results from a global top-down form recognition process, rather than from a bottom-up local motion analysis. This conclusion was challenged by Giese and Poggio (2003), who proposed that a hierarchical bottom-up process using only local motion signals combined with an attention process could account for the results. Neri, Morrone, and Burr (1998) claimed that the perception of biological motion in the presence of noise is driven mainly by the integration of local motion signals.

Studies that emphasized the contribution of local motion signals often argue that the information from a

single static picture of a point-light walker does not allow a naive observer to perceive a walking human figure. Spontaneous biological motion perception occurs only in an animated sequence (Johansson, 1973). Therefore, most studies on biological motion perception have suggested or implicitly relied upon the assumption that the perception is processed by means of local image motion signals (Cutting, 1981; Johansson, 1973; Mather, Radford, & West, 1992; Neri et al., 1998). However, while a single static frame is insufficient to recognize a walker, biological motion perception might also be derived from temporal integration of the sparse form information in each frame.

Computational studies have also emphasized the role of local motion signals. Giese and Poggio (2003) proposed a model that analyzed form and motion cues separately. Their model accounts for a variety of experimental results purely by using the extracted local motion signals. In contrast, the form-analyzing pathway did not reveal selectivity for biological motion stimuli. Based on Giese and Poggio's approach, Casile and Giese (2005) developed a model that relied on the local motion signals in the stimulus. This model contained detectors of local motion signals that move in opposing direction. Casile and Giese computed the amount of "opponent motion" signals in the stimulus proposed by Beintema and Lappe (2002) and developed a new artificial stimulus with the same amount of opponent motion signals. From the approximate similarities between the two stimuli and the corresponding model simulations, Casile and Giese claimed that these opposing local motion signals might act as a critical feature in biological motion perception. This debate clearly reveals the controversy relating to which processes are necessary for perceiving biological motion as opposed to those supplementary in nature.

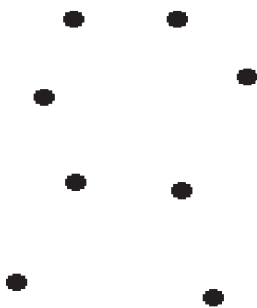
While several studies investigated the contribution of local motion signals, our objective in this study was to investigate quantitatively the contribution of global form information. We present a simple model based on

template matching, which relies on form analysis only and completely ignores any image motion signals. We investigated how much form information is available from point-light walkers and whether this information could contribute to tasks that use point-light walkers as a stimulus. By comparing the performance of the model to both the psychophysical results described above and to the additional experiments reported below, we assessed quantitatively the contribution of form information. Among the many and often complicated characteristics of biological motion, we will focus on basic and often used low-level discrimination tasks. We chose these tasks on the one hand because they are simple and allow a straightforward quantitative comparison and, on the other hand, because we believe that restricting the scope of the model is advisable for an early investigation. For a similar reason, we concentrated on stimuli without masking noise. Beintema and Lappe (2002) have argued that biological motion recognition within noise may involve not only the perception of the biological motion stimulus per se but also the segmentation of the figure from the background, which could be a different process. The relationship between our model and the masking studies will be considered in the [Discussion](#) section.

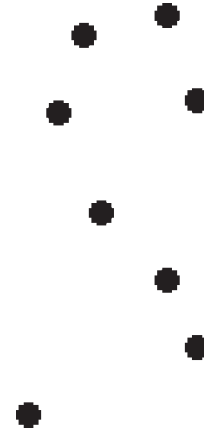
Methods

Stimuli

Stimuli were computer-generated two-dimensional point-light walkers (Cutting, 1978). All translatory movement components were eliminated, giving the impression of a person walking on a treadmill. Each trial simulated a walking speed of 1.6 s per gait cycle, with each stimulus frame presented for 52 ms. The trials lasted for one gait cycle (31 frames), except for the forward/backward task in the first experiment ([Figure 3c](#)), which lasted 2 s (i.e., 40 frames, walking speed 1.6 s per gait cycle), in analogy to the experimental parameters (Beintema et al., [in press](#)).



Movie 1. Movie of the point-light walker generated by Cutting's (1978) algorithm. Click on the image to view the movie.



Movie 2. Movie of the new stimulus developed by Beintema and Lappe (2002). Click on the image to view the movie.

Three different types of point-light walkers were used. The first type is the classical walker introduced by Johansson (1973), in which point lights appear on the major joints of the body and produce smooth trajectories when the stimulus is in motion ([Movie 1](#)). The second type, introduced by Beintema and Lappe (2002), manipulated these stimuli such that the single dots did not keep a constant position on the body but rather changed position each frame by jumping to a new, randomly selected position on the limbs ([Movie 2](#)). This minimized local motion signals and allowed to selectively manipulate them by varying the lifetime of the dots (in number of frames) before a new position is allocated. Note that because the dot positions are chosen randomly each frame, any single trial comprises different stimulus frames than any other trial.

The third type of stimulus was introduced by Casile and Giese (2005). It is similar to the stimulus used by Beintema and Lappe in terms of local motion signals, but it is degraded in terms of positional information. The stimulus consisted of four regions. In two of these regions, roughly corresponding to the position of hands and feet, dots move with a sinusoidal horizontal component and with a random vertical component. The other two regions contain dots that move completely random. The spatial arrangement of the four regions was derived from the spatial arrangement of a person walking to the right or to the left.

Tasks

Following the psychophysical studies to be simulated, we used three different tasks to compare the results of the template-matching analysis with psychophysical data.

Direction task

In this task, human observers and the template-matching model had to decide whether the walker was facing and moving to the right or to the left.

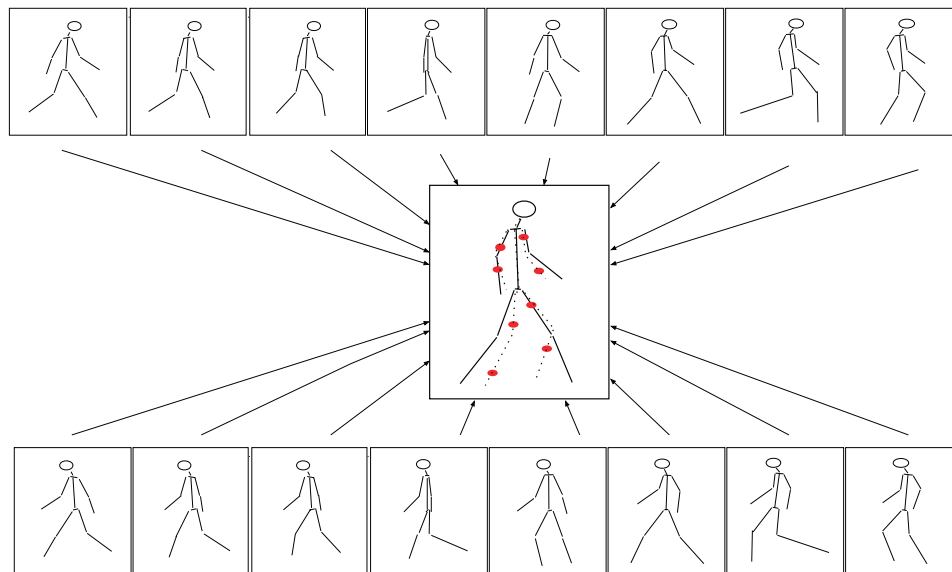


Figure 2. One frame of the stimulus (filled circles; the dashed lines are only for clarification and are not shown in the real stimulus), which is matched to a set of templates of a walker moving and facing to the right and a walker moving and facing to the left (stick figure, solid lines). The match depends on distance measurements between stimulus dots and the template.

Coherence task

Here, the template-matching model and the human observers had to decide whether the upper and lower parts of the body were facing and moving in the same (coherent) or opposite (incoherent) direction. This task was used previously by Mather et al. (1992). For the model, this decision was similar to the direction task except for the fact that the upper and lower halves of the stimulus were initially treated separately. After making a direction decision as described for the direction task for each half separately, the model decided whether both halves were walking in the same or in different directions.

Forward/backward task

In this task, the model and the human observers had to decide whether the walker was moving in a forward or backward direction (Beintema et al., *in press*). Both conditions comprised identical frames. The sequence of frames was shown either in correct order, giving the impression of a walker moving forward or in reversed order. In this case, the walker appeared to move backward.

Templates

We attached sensors to the major joints of the bodies (i.e., shoulders, elbows, wrists, hips, knees, and ankles) of nine people (five were male) and recorded their walking movements on a catwalk with a body-tracking system (MotionStar, Ascension). Because we used only two-dimensional stimuli, we omitted the depth component of

the movement patterns. We spatiotemporally averaged (Giese & Poggio, 2000) the individual walking patterns and connected the dots in a biological appropriate way, resulting in a stick figure model of a “mean walker.” We subdivided one step cycle of this walking pattern into 100 temporally equidistant frames. The model used these frames as its templates of a common walking person (Figure 2). The number of movement patterns that were used to generate the templates is arbitrary. We assured that the number is high enough to avoid much influence of an individual person’s movement pattern and, thus, to ensure that the templates represent an arbitrary average of human locomotion patterns. Similarly, the number of templates was set to 100 because the number approximately matched the temporal resolution of the tracking system and because the temporal sampling of the walking sequence seems reasonable. In informal tests, we confirmed that varying the number of templates between 50 and 150 did not affect the performance of the model in the task we present below.

Template-matching analysis

The template-matching analysis was achieved by a frame-by-frame template-matching algorithm, which evaluates the distances between the templates and the stimulus frames (Figure 2).

For each stimulus frame, the model computed the shortest Euclidian distance of every dot in the stimulus frame to any of the limbs for each possible template. Thereby, stimulus dots were not restricted to a specific limb nor was the number of dots per limb restricted. Out of all computations, the model selected the shortest distance. After summing

these minimum Euclidian distances of all dots in each frame, the frame with the shortest total distance was selected from the set of template frames. This choice was based on a winner-take-all principle. In preliminary tests, we checked the dependence of the model on the distance measure. Instead of the linear Euclidian distance measure, we used variants like quadratic, cubic, or Gaussian distance functions for the discrimination task. The differences for the recognition rates between these and the simple Euclidian distance were less than 4%. Thus, in the simulations reported below, we used the simple and parameter-free Euclidian measure.

By matching each stimulus frame to 100 template frames of a walker moving to the right and to 100 template frames of a walker moving to the left, the model had to decide frame by frame which direction decision it would favor. When one whole stimulus was completed, all single decisions were averaged to achieve a decision in the respective task. In a subsequent processing stage, the model analyzes the temporal position of the best matching template in the whole set of templates. It computes whether consecutive frames were arranged in the order expected when the walker was moving forward or backward. In this manner, the model constructed chains of consecutively ascending or descending frames. At the end of one trial, the chain with the maximum length was used as the decision variable. A consecutively ascending temporal order of the selected frames was an indication for forward movement; a consecutively descending temporal order of the frames was an indication for backward movement.

Thus, the model only uses the available form information, ignoring all motion signals, if there were any (see [Appendix](#) for mathematical description).

In each simulation run for a given task, we presented 100 trials. Although we always used the same stimulus, trials differed because the single dots of the stimulus were always drawn on different positions. A full step cycle of a walking sequence was divided into 31 frames. In the forward/backward simulations, a single trial lasted for 2.0 s and consisted of 40 frames (thereby keeping walking speed constant at 1.6 s per gait cycle), in accordance with the parameters used in the psychophysical study of that task (Beintema et al., [in press](#)). The model calculated decisions for each trial as described above. We then summed the 100 single decisions for each of the 100 trials to calculate the overall recognition rate for the task. During the simulations, all stimulus properties like trial duration and stimulus size were identical to the conditions used in the psychophysical tasks. Starting phase of the walking cycle of the stimulus was randomized over trials.

In comparing psychophysical with computational data, we needed to account for the phenomenon of visible persistence (Coltheart, 1980). Visible persistence refers to the fact that light points presented to an observer for a period shorter than 100 ms are perceived for as long as 100 ms, whereas dots shown for longer periods are perceived for the time they are actually presented.

In psychophysical experiments, subjects reported seeing more points on the screen than were presented in any single frame. Quantitative analysis showed that in accordance with the literature reviewed by Coltheart, subjects perceived about twice as much dots than are really shown at the 50-ms frame duration (Beintema et al., [in press](#)). We, therefore, feel that visible persistence is part of the process of interpreting these stimuli and consequently needs to be implemented in the template-matching analysis. We adapted the model to this effect as simple as possible: to include the effect of visible persistence by overlapping the dots in a stimulus frame with the dots of the preceding frame if the presentation duration of the frame was less than 100 ms. The resulting frame then had twice the number of dots: the dots from the frame itself plus the dots from the previous frame.

The model uses a view-based approach that treats size and position of the stimulus as constant. We believe that these assumptions, especially knowledge of height and position, are appropriate for a template-matching model when a discrimination stimulus is presented in isolation as in the experiments that we modeled. The model does not use any adjustable parameters that could be fitted to the psychophysical data. The model stages were chosen to be as simple and intuitive as possible.

In the simulations, we compared the model's decisions in each processing stage with the stimulus properties and determined the percentage of correct decisions within 100 trials, each containing a full walking cycle of the stimulus. Note that the stimulus is a computer-generated artificial walker, whereas the templates were obtained from recordings of actual human walkers. Therefore, the stimulus will never exactly match any of the templates. Thus, the model is not expected to yield recognition rates of 100%. We believe that this is an appropriate comparison with the psychophysical task in which this same computer-generated walker was presented to human observers. If, as we predict, human observers use templates of body postures, then it is likely that these templates are also learned from observing real people walking. We compared the recognition rates of the model to data from psychophysical experiments or psychophysical data obtained from other studies with the same tasks and stimuli as used in the model simulations.

Results

We intended to test the performance of the model in several tasks, in which we varied parameters that influence form and motion contribution as well as stimulus types.

First, we studied the importance of form information. We used the psychophysical data from the studies by Beintema and Lappe (2002) and Beintema et al. ([in press](#)) and simulated these tasks with our model. We varied the amount of form information by presenting different

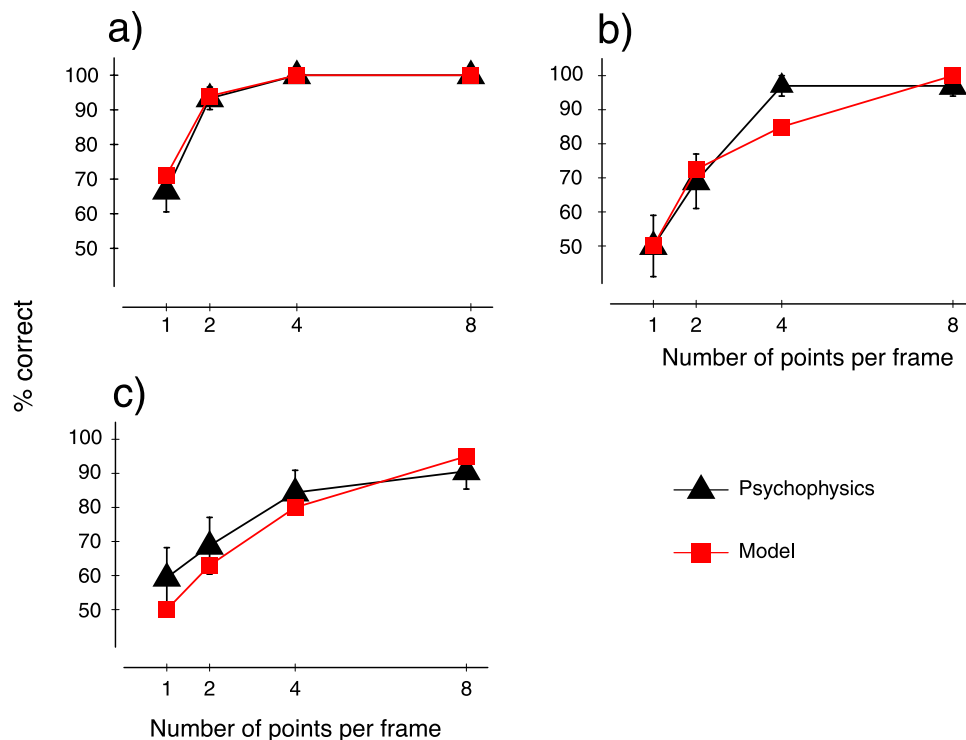


Figure 3. Percentage of correct answers as a function of the number of points shown per frame for (a) the direction task, (b) the coherence task, and (c) the forward/backward task for model and human observers. Psychophysical data are adapted from Beintema et al. (in press) and Beintema and Lappe (2002) and are shown as mean \pm SE.

numbers of dots per frame (1, 2, 4, or 8). The frame duration and the lifetime of the dots were kept constant (52 ms = 1 frame).

We first simulated a direction discrimination task and compared the results to data from Beintema and Lappe (2002; for the results, see Figure 3a). We secondly simulated a coherence task and compared the results to data from Beintema et al. (in press; for the results, see Figure 3b). In both tasks, we presented one step cycle of the stimulus (31 frames) consistent with the experimental studies.

The model was able to solve both the direction and coherence tasks solely based on static form information. Therefore, both tasks can be solved in principle from form analysis alone and do not necessarily depend on the perception of the temporal pattern of walking. In the forward/backward task introduced by Beintema et al. (in press), recognition of the stimulus depends on temporal integration. Both stimuli (forward and backward walking) consisted of the same individual frames; the only difference was the order in which they were shown. In normal frame order, the impression of a forward-moving walker occurred, whereas in reversed order, the impression of a backward-moving walker occurred. We simulated this task with the second stage of the model, which analyzes the temporal order of the frames and compared the results to data from Beintema et al. (for the results, see Figure 3c). Beintema et al. used a stimulus duration of 2 s while keeping the walking speed of the stimulus constant.

Therefore, we simulated this task with a stimulus duration of 40 frames (one gait cycle still comprises 31 frames).

We statistically analyzed data from Beintema and Lappe (2002) and Beintema et al. (in press; Figure 3). A two-way analysis of variance (ANOVA) with “number of points” as within-subject factor and “task” as between subject factor revealed a significant main effect of number of points, $F(3,12) = 18.27$, $p < .01$. No statistically significant effects were found on the factor task, $F(2,12) = 3.8$, $p = .12$, and on the interaction between task and number of points, $F(2,12) = 0.66$, $p = .57$. Planned within-subjects contrasts revealed a significant linear trend, $F(1) = 26.65$, $p < .01$, stating that recognition rates increase with increasing number of points. Three subjects participated in the study of Beintema and Lappe (Figure 3a), and there were two subjects in the study of Beintema et al. (in press; Figure 3b and 3c).

As shown in Figure 3, the results of the model simulated the human performance accurately. To quantify the model simulations, we compared the model data and the mean value of the psychophysical data for each data point separately. No statistical differences could be observed (one-sample t tests, $p > .05$).

Local motion signals

We next looked at the influence of local motion signals on the perception of biological motion. To compare model

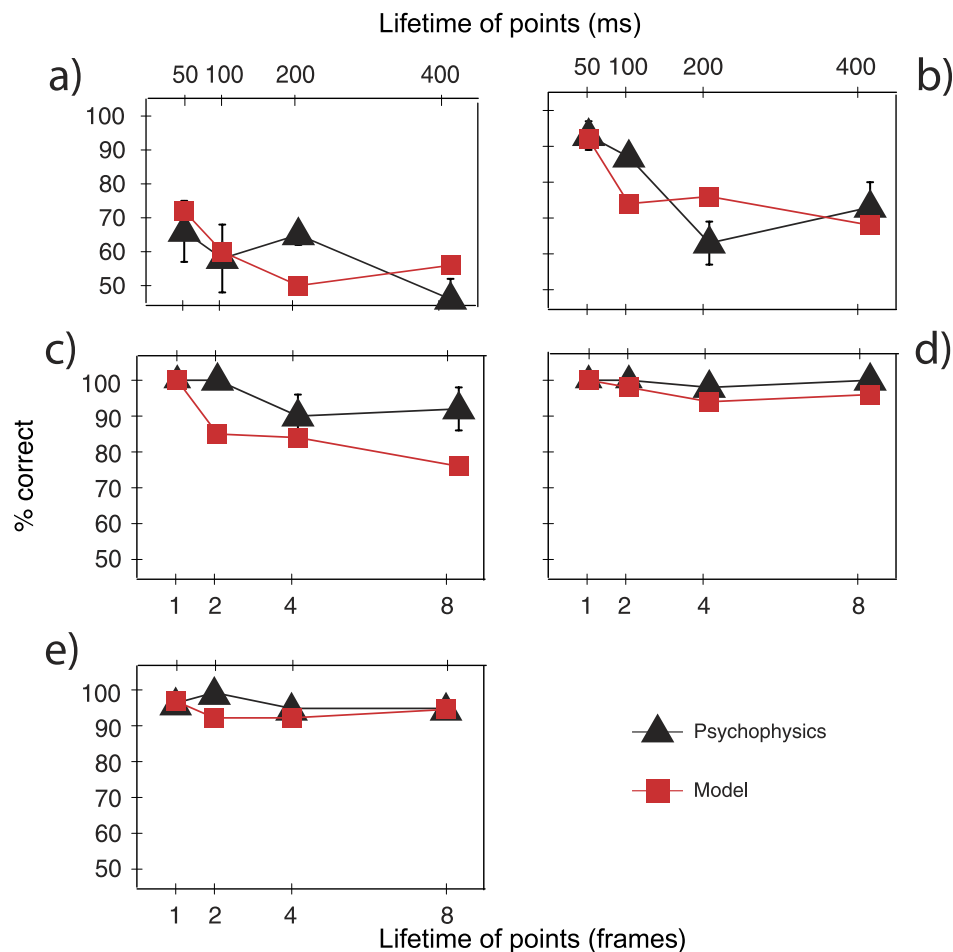


Figure 4. Percentage of correct answers as a function of the lifetime of single points in the direction task for (a) 1 point per frame, (b) 2 points per frame, (c) 4 points per frame, and (d) 8 points per frame and in the (e) forward/backward task for 8 points per frame. Psychophysical data (Panels a–d were adapted from Beintema & Lappe, 2002; Panel e was adapted from Beintema et al., *in press*) are shown as mean \pm SE.

simulations with psychophysical data, we adapted the data of Beintema and Lappe (2002) and Beintema et al. *in press*. Beintema and Lappe added local motion signals to the stimulus by varying the duration (number of frames) for which a dot kept its position on the limb before being extinguished and relocated to a different position. If a dot remained at a specific limb position for several frames, its spatiotemporal profile allows estimating local image motion. Because the model, however, relies on form information only, it does not evaluate this local motion signal. If human observers do take advantage of local motion signals, the answers of the model and humans should therefore differ. The difference should become more obvious with prolonged lifetime of the dots.

The simulations were conducted with the direction task (with 1, 2, 4, and 8 points per frame) to examine whether local motion signals would improve performance in general (Figure 4a–d), as well as with the forward/backward task (8 points per frame), in which local motion

signals should provide the most useful additional information (see Figure 4e). In addition to varying the lifetime of the dots, we also varied the number of dots per frame. Human data in the direction task were taken from Beintema and Lappe (2002). Data in the forward/backward task were taken from Beintema et al. (*in press*).

Prolonging lifetime results in an increase of local motion signals. Therefore, an increase of correct answers with prolonging lifetime should be expected if the perception relies on local motion signals. We statistically analyzed the data from Beintema and Lappe (2002; Figure 4a–d). A two-way ANOVA with “lifetime of points” as within-subject factor and number of points as between-subject factor revealed a significant main effect of lifetime of points, $F(3,24) = 4.74$, $p = .01$. There was also a statistically significant effect on the factor number of points, $F(3,8) = 60.37$, $p < .01$, and no effect on the interaction between lifetime of points and number of points, $F(3,8) = 1.1$, $p = .40$. Planned within-subjects contrasts

revealed a significant linear trend, $F(1) = 6.72$, $p = .03$, stating that recognition rates decrease with increasing lifetime of points. Three subjects participated in each experiment.

The observed behavior would be expected if form information were dominant over local motion signals. With longer lifetime, each dot remains on a fixed position on the body for a longer time. Thus, the sampling rate of the form of the body is reduced. Also, the number of dots effectively perceived due to visible persistence decreases when lifetime is prolonged. In the model, this is equivalent to decreasing the number of effectively used dots. For this reason, the performance of the model drops if lifetime alters from one to two frames. For the other transitions (from two to four and from four to eight frames), the recognition rates are constant and differences are negligible because they show no uniform trend. We observe a similar behavior for the psychophysical data. An a priori contrast test between the conditions “lifetime” of one frame and two frames revealed a significant effect, $F(1,8) = 16.0$, $p < .01$, whereas the contrast test between two frames and four frames and that between four and eight frames no longer show a significant effect, $F(1,8) = 4.46$, $p = .07$ and $F(1,8) = 0.12$, $p = .74$, respectively. Furthermore, we found a statistically significant difference between model and psychophysical data only for two conditions: one point per frame, lifetime four frames and two points per frame, lifetime two frames (both: one-sample t test, $p < .05$). Because of ceiling effects, for the condition four points per frame, lifetime two frames, no t value could be calculated.

For the forward/backward task (Figure 4e), there was no significant influence of lifetime, ANOVA with repeated measures: $F(3,6) = 3.17$, $p = .11$. Similarly, the model does not reveal a marked influence on lifetime in the forward/backward task. Furthermore, there were no statistically significant differences between model data and psychophysical data (one-sample t test, $p > .05$).

In summary, the model matched the performance of human subjects qualitatively and quantitatively, although it uses form information only and ignores local motion signals.

Other walkers

We tested discrimination tasks also with the classical point-light walker introduced by Johansson (1973).

Figure 5 shows the results of the simulation of a psychophysical experiment by Mather et al. (1992). They presented the classical walker in a direction discrimination task in conditions in which specific dots were omitted from the walker. In the first condition, all dots were shown, whereas in the other conditions, four dots were removed in any of the following: shoulders and hips, elbows and knees, or wrists and ankles. Particularly, the omission of wrists and ankles had a deteriorating effect on perception.

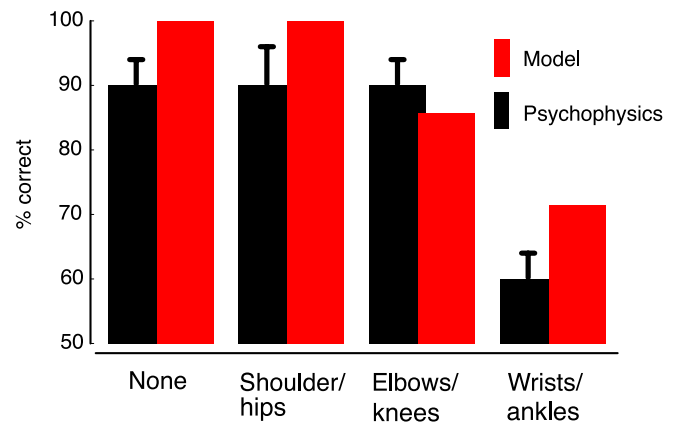


Figure 5. Percentage of correct answers as a function of the dots omitted. Psychophysical data (adapted from Mather et al., 1992) are shown as mean \pm SE.

The results of the template-matching analysis revealed the same dependence on visible dots as the psychophysical data of Mather et al. Leaving out the ankles and wrists impaired the perception. Leaving out the elbows and knees showed little effect, whereas omitting the shoulders and hips had no influence at all.

Mather et al. concluded that the feet and hands are most important because they follow the longest trajectories, hence providing most motion information. Our simulations revealed similar results and confirm that distal dots are most important. Because the model uses only form information, we conclude that the reason that wrists and ankles are more important than other, more proximal dots is that they offer the most reliable spatial information about the posture of the walker.

Casile and Giese (2005) argued against the idea that the form information in the point-light walkers with strongly degraded local motion information is sufficient to explain psychophysical data by Beintema and Lappe (2002). Casile and Giese proposed a model that used only “opponent local motion” features to model the psychophysical data of Beintema and Lappe (2002) and for a newly developed stimulus.

The new stimulus (critical feature stimulus [CFS]) proposed by Casile and Giese was similar to the stimulus used by Beintema and Lappe in terms of local motion signals, but it was degraded in terms of positional information. The stimulus consisted of four regions. In two of these regions, roughly corresponding to the position of hands and feet, dots move with a sinusoidal horizontal component and with a random vertical component. The other two regions contain dots that move completely random. The spatial arrangement of the four regions was derived from the spatial arrangement of a person walking to the right or to the left. Casile and Giese’s psychophysical data with the CFS are similar to the psychophysical results by Beintema and Lappe (2002).

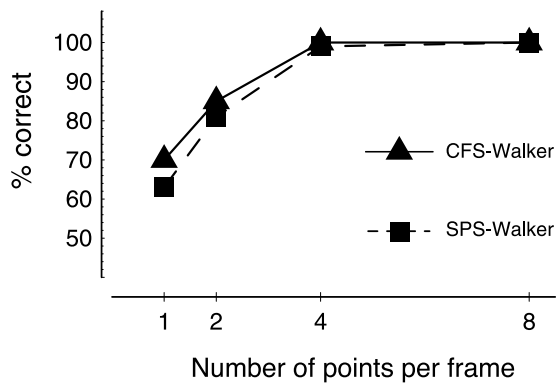


Figure 6. Comparison of the template-matching analysis in a direction task for the stimulus used in this study (SPS) and the stimulus proposed by Casile and Giese (2005; CFS-Walker).

Beintema and Lappe reported a counterintuitive decline of recognition rates for prolonged lifetimes of the stimulus dots (Figure 4). Although Casile and Giese were unable to replicate this decline of recognition rates, they claimed that the remaining sparse local image motion information acted as a critical feature for the recognition.

We used this CFS-Walker to test the available form information in the CFS-Walker stimulus with our template-matching analysis. Recognition rates for the CFS-Walker in our template-matching model were similar to those for the sequential position stimulus (SPS)-Walker. Both stimuli revealed the same dependency on points per frame (Figure 6). These simulations showed that the positional information in the CFS-Walker is still comparable with the information in the SPS-Walker and that this information is sufficient to explain the psychophysical data.

As described above, the walker consisted of four regions, each with a specific spatial offset from the vertical axes. If there would be no spatial displacement of the four regions at all, a stimulus facing to the left would be identical to a stimulus facing to the right. The discrimination task would be unsolvable, and thus, the spatial displacements are essential to observe results different from chance level. Therefore, it is these spatial displacements that act as the critical feature rather than the opposing motion signals.

Although the CFS-Walker matched the form of a human body only very roughly, the performance of this model was similar to that of the SPS-Walker. At first glance, this result seems surprising because we compare a walking stimulus to an artificial stimulus with coarse position information approximately matching the original stimulus. However, in the SPS-Walker, which is mainly used in this study, the stimulus dots also almost never exactly match the template. Actually, the SPS-Walker also represents an artificial computer stimulus. Therefore, the model identified the SPS-Walker, as well as the CFS-Walker, as a noisy stimulus of a walker. The simulation results ob-

tained with both the CFS-Walker and the psychophysical results by Casile and Giese indicate that the visual system is very robust against noise.

Stimulus duration

In addition to already existing psychophysical studies (Beintema et al., *in press*; Beintema & Lappe, 2002; Casile & Giese, 2005; Mather et al., 1992), we conducted a further experiment and compared the data to simulation results.

Methods

The new experiment used a direction task and varied the duration of the stimulus. The methods followed the procedures described in Beintema et al. (*in press*) and Beintema and Lappe (2002). The stimuli were presented on a monitor with a resolution of 1280×1024 pixels and a display size of 30×40 cm. The duration of a single frame was 52 ms. The lifetime of each single dot was 1 frame, that is, 52 ms. The subjects were seated 60 cm in front of the monitor and viewed the stimulus binocularly. The stimulus covered a field of 5×10 deg and consisted of white dots (5×5 pixels) on a black background. Stimulus position had a randomly chosen offset. Stimulus duration was varied from about 100 ms to 1.6 s (2 to 31 frames) in a pseudorandomized manner. In blocked trials, 2, 4, or 8 stimulus dots per frame were presented. The subjects had to decide the direction of the walker after each trial and indicate their decision with a button press, whereupon the next trial started. Six subjects participated.

Results

The results are shown in Figure 7. A two-way ANOVA with “stimulus duration” as within-subject factor and number of points as between-subject factor revealed a significant main effect of stimulus duration, $F(4,60) = 54.22$, $p < .01$. Also, the factor number of points revealed a significant effect, $F(2,15) = 39.7$, $p < .01$, as well as the interaction between number of points and stimulus duration, $F(2,15) = 13.0$, $p < .01$. Planned within-subjects contrasts revealed a significant linear trend, $F(1) = 168.0$, $p < .01$, stating that recognition rates increase with prolonged stimulus duration.

The model revealed the same qualitative behavior; recognition rates increased with prolonged lifetime and with increasing number of points per stimulus frame. For sparse form information, that is, for 2 and for 4 points per frame and very short stimulus durations (100 and 200 ms), however, the model significantly overrates the recognition rates (one-sample *t* test, $p < .05$). Consequently, the model in these simulations was performing better than the average of the human observers. This might be caused by two reasons: For simplicity, the template-matching

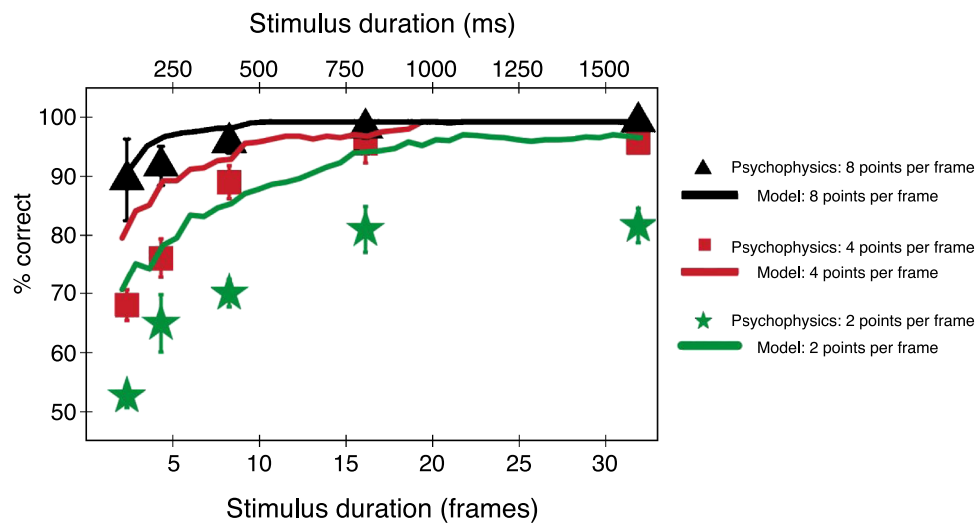


Figure 7. Percentage of correct answers as a function of the total stimulus duration for 2, 4, and 8 points per frame in the direction task. Psychophysical data are shown as mean \pm SE.

analysis is not affected by internal noise, whereas humans apparently are affected. Especially for sparse visual input, the neural noise in the human visual system will presumably attenuate humans' ability to correctly solve the task. Secondly, the human subjects in this experiment were untrained to biological motion perception. By contrast, the model must be considered as a trained observer because it is able to extract the form information accurately. Trained human observers, like those in the experiment by Beintema and Lappe (2002) presented above, are able to reach higher recognition rates comparable with the recognition rates of the template-matching analysis even for sparse form information (Figure 3).

Discussion

In this article, we addressed the question of possible mechanisms underlying the perception of biological motion. Global biological motion, that is, the motion of the human figure, may be derived from local image motion analysis of the light points or from structural information from the changing shape of the body.

In our study, we investigated quantitatively the contribution of global form to discrimination tasks with point-light walkers. By assuming a library of static postures of a walking person, we developed a model based on template matching, which uses only sequential posture information rather than local motion signals. In this way, we could quantify the amount of global form signals available in the depicted point-light displays in the absence of local image motion signals. Global motion information will be derived from the change of these postures over time rather than from local motion signals.

First, we employed different experiments that varied the amount of form and local motion signals and compared the results of the template-matching analysis to the psychophysical data. Data in some experiments were taken from previous studies (Beintema et al., *in press*; Beintema and Lappe, 2002; Figures 3 and 4). One experiment was a new one conducted for this study (Figure 7). In these experiments, we varied the number of visible points per frame, the stimulus duration, and the amount of local motion signals. The comparison revealed a strong dependence on available form information in the template-matching model similar to the data from the psychophysical studies. These similarities were consistent when the total form information was not varied not within one single frame but in the overall information mediated by stimulus duration.

Beintema and Lappe also reported a counterintuitive decline of recognition rates for prolonged lifetime of the stimulus dots. Casile and Giese (2005) proposed a model that was supposed to explain these psychophysical data by exploiting local motion features. However, the model predicted a slight increase of recognition rates for a prolonged lifetime of stimulus dots. In contrast to Casile and Giese, our simulation results based on global form showed a decline for a longer lifetime of stimulus dots similar to the psychophysical data. In accordance with Beintema et al. (*in press*), we suggest that this decline is due to a decreasing form sampling in combination with visible persistence and not due to the addition of local motion signals. For a detailed analysis of the relationship of recognition rates and visible persistence, see the study of Beintema et al.

In another experiment, Casile and Giese introduced a new artificial stimulus called CFS. They intended to show that this stimulus does not contain any information

but only opponent local motion features. They claimed these opposing motion vectors to be critical features that are essential to recognize biological motion. Our simulations showed that the CFS-Walker still contained coarse global form information. This information is strongly degraded but is still sufficient to solve the applied task (Figure 6). We suggest that the critical feature in this stimulus is global form and not local motion signals.

The data on the CFS as well as on the study by Mather et al. (1992) also demonstrate that the results presented in this study do not depend on the particular stimulus used. Our template-matching analysis was able to simulate data from the classical Johansson walker as well as from the stimuli of Beintema et al. and Casile and Giese. Furthermore, the simulations on the study by Mather et al. suggest that the results can be explained by missing form information instead of missing motion information (Figure 5, but see also Troje & Westhoff, 2006).

Several studies examined the perception of biological motion in masking experiments; that is, the stimulus is shown in a number of distracting noise dots. Beintema and Lappe (2002) have argued that these tasks comprise both the perception of the biological motion stimulus per se and the segmentation of the stimulus from the background, which could be a separate process. Within the current analysis, we did not simulate noise experiments because we wanted to keep the model simple and confined to the biological motion task without an additional segmentation stage. Such a separate segmentation stage, however, might not be needed. Lee and Wong (2004) recently presented a template model for the recognition of biological motion that is similar in spirit to our approach but uses point-light templates rather than stick-figure templates. They showed that a form-based template-matching approach could also account for perception of biological motion in noise with results similar to psychophysical data (Neri et al., 1998). In addition, because their model did not include segmentation by local image motion, it should also work for more complicated noise patterns. Although more work would be needed to confirm this hypothesis for our approach, we suggest that our template matching would be able to reveal similar results.

Our computational approach suggests that perception of biological motion is possible from matching the sparse stimulus frames to (dynamic) form templates and integrating this information over time. The concept that learned global prototypes underlie the interpretation of perceived body structures in a top-down process relates back to initial ideas by Marr and Nishihara (1978). The general idea is supported by other psychophysical experiments. Sinha and Poggio (1996) connected the dots on the major joints of a human body so that it showed the line drawing of a person. If this rigid structure was rotated about its vertical axis, it was seen as

a walking figure. However, if the wrong joints were connected (so that the figure did not represent a human body), the rotation was correctly interpreted. Sinha and Poggio argued that the visual system interprets the human figure in terms of how the human structure is expected to change. Our study supports this idea of a form-based top-down process mediating the perception of biological motion. The simulations reveal that Sinha and Poggio's idea can be extended from static line drawings to moving point-light figures. Even if a single frame did not provide enough information to recognize the human figure, the succession of point-light images was sufficient. The simulations do not exclude the fact that human observers can use available local image motion signals if they are useful.

Our approach derives global motion information from an analysis of the changing shape of the figure rather than from local motion detectors. As such, it bears some relationship to feature-based motion systems as suggested by Cavanagh (1992) and Lu and Sperling (1995). From a computational viewpoint, the advantage of such feature-based motion systems over lower level, energy-based motion systems is particularly high for biological motion recognition because in contrast to rigid object motion, biological motion relies inevitably on form information due to the large number of degrees of freedom in the nonrigid motion of the body.

Perception of biological motion is not just a single phenomenon. The perception of biological motion is composed of a rich palette of different aspects such as action recognition (Dittrich, 1993, Johansson, 1973; Pollick, Fidopiastis, & Braden, 2001), gender discrimination (Pollick, Lestou, Ryu, & Cho, 2002; Troje, 2002; Troje, Westhoff, & Lavrov, 2005), and identification of identity (Cutting & Kozlowski, 1977; Loula, Prasad, Harber, & Shiffrar, 2005). It has been shown that humans can use different information to judge movements depending on the task (Pollick et al., 2001, Troje, 2002) and that the influence of bottom-up and top-down processing, as well as attention, differs among tasks (Thornton, Pinto, & Shiffrar, 1998; Thornton, Rensink, & Shiffrar, 2002; Thornton & Vuong, 2004). In this study, we have focused on straightforward discrimination tasks for simplicity. These simple tasks can be solved by a global form analysis in the absence of local motion signals. One may now ask: Which of the more complex aspects of biological motion perception are local and which ones are global? Which ones require motion per se and which ones are based on structural cues? In principle, a template model such as ours may be sufficient to also discriminate action, gender, or identity provided that the appropriate templates are available. The model arrives at a description of the temporal structure of the body posture change over time and, thus, may also discriminate actions and use dynamic cues (Troje, 2002) even if they are not derived from local motion analysis. Whether this is truly sufficient would have to be investigated in further studies; however,

it is also likely that among the many aspects of biological motion, there are some that benefit from additional motion signals. However, for the task we studied here, these local motion signals do not form a critical feature for biological motion.

Appendix

This section provides a mathematical description of the model. The model consists of two hierarchically arranged but functionally separate stages.

In Stage 1, the model uses a library of size-normalized static templates T with known two-dimensional coordinates x_p^T for each stimulus point p . The set of template points x_p^T comprises not only the joint positions of the template but also all positions on the lines connecting the joints (i.e., $p \rightarrow \infty$). The input to Stage 1 are the coordinates x_i^S of the stimulus dots i of a given frame S . The model computes the distances $d_L^{S,T}$ and $d_R^{S,T}$ between a given stimulus frame S and each of the templates T_L (templates for walking to the left) and T_R (templates for walking to the right) by calculating the minimum Euclidian distance between each of the stimulus dots x_i^S and all locations x_p^T on each template frame and adding all single distances up. Technically, the distance was measured by dropping a perpendicular line on the nearest limb or by calculating the distance to the nearest joint, whichever was shorter. This procedure was performed independently for each set of templates T_L and T_R without adding any internal or external noise:

$$d^{S,T_L} = \sum_{i=1}^n \min_p (|x_i^S - x_p^{T_L}|),$$

$$d^{S,T_R} = \sum_{i=1}^n \min_p (|x_i^S - x_p^{T_R}|),$$

where n is the number of stimulus dots.

For a given stimulus frame S , the best matching templates were determined by finding the templates with the minimum distances d^{S,T_L} and d^{S,T_R} independently for each template set T_L and T_R :

$$d_{\min}^{S,T_L} = \min_{T_L} (d^{S,T_L}) = d^{S,T_L^{S,\min}},$$

$$d_{\min}^{S,T_R} = \min_{T_R} (d^{S,T_R}) = d^{S,T_R^{S,\min}}.$$

This procedure results in $T_L^{S,\min}$ and $T_R^{S,\min}$, which represent the template frames within the template set for walking to the left (T_L) and for walking to the right (T_R) that match the stimulus frame S best. $d^{S,T_L^{S,\min}}$ denotes the distance between the given stimulus frame S and the best matching template frame $T_L^{S,\min}$ for walking to

the left, and $d^{S,T_R^{S,\min}}$ denotes the distance between the given stimulus frame S and the best matching template frame $T_R^{S,\min}$ for walking to the right.

The model's decision variable at Stage 1 (c_s^1) to discriminate the walking direction of the stimulus in a single stimulus frame S is based on the minimum distance measures $d_{\min}^{S,T_{L,R}}$:

$$c_s^1 = 1 \text{ for } d_{\min}^{S,T_L} < d_{\min}^{S,T_R} \text{ and } c_s^1 = -1 \text{ otherwise.}$$

For $(c_s^1)^1 = 1$, the model decides in favor of walking to the left; for $(c_s^1)^1 = -1$, the model decides in favor of walking to the right. Note that the templates for left and right (T_L and T_R) are never identical, and therefore, $d_{\min}^{S,T_L} = d_{\min}^{S,T_R}$ is never the case.

A trial consists of N stimulus frames. Each frame is evaluated independently from the other frames by the computation described above. At the end of a given trial, the model computes an overall decision variable at Stage 1 c^1 by averaging all single decision variables $(c_s^1)^1$:

$$c^1 = \frac{\sum_{s=1}^N c_s^1}{N}.$$

For $c^1 > 0$, the model decides in favor of walking to the left; for $c^1 < 0$, it decides in favor of walking to the right. For the rare case of $c^1 = 0$, the model randomly decides in favor of left or right.

This procedure is applied to each of the 100 trials of a simulation run, and the proportion of correct decisions is expressed as the percentage of correct decisions.

The best matching templates $T_L^{S,\min}$ and $T_R^{S,\min}$ for all stimulus frames S are forwarded to model Stage 2. In this stage, the template frames are ordered depending on their temporal position in the entire walking sequence from 1 to t . For two consecutive stimulus frames S and $S + 1$, the decision variable in Stage 2 ($c_{S,S+1}^2$) is:

$$c_{S,S+1}^2 = 1 \text{ for } T_{L,R}^{S,\min} \leq T_{L,R}^{S+1,\min} \text{ and}$$

$$c_{S,S+1}^2 = -1 \text{ for } T_{L,R}^{S,\min} \geq T_{L,R}^{S+1,\min}.$$

If, for two consecutive stimulus frames S and $S + 1$, the best matching template frames are from the same template set (i.e., both frames are out of the set T_L or both are out of T_R) and these two template frames are temporally ascending, then the variable is set to $c_{S,S+1}^2 = 1$. If the selected template frames are descending, $c_{S,S+1}^2 = -1$. If the selected frames for S and $S + 1$ are equal, $c_{S,S+1}^2$ is set equal to the preceding entry for $c_{S,S+1}^2$. In case the selected template frames for the stimulus frames S and $S + 1$ are from different template sets (e.g., for S from T_L and for $S + 1$ from T_R), the variable is set to $c_{S,S+1}^2 = 0$.

An entire trial consists of N stimulus frames. This leads to a time series T with $N - 1$ entries for $c_{S,S+1}^2$. An overall decision variable after one trial for a forward (c_f^2) and a backward (c_b^2) movement is achieved by applying two functions F_f and F_b on the series T :

$$c_f^2 = F_f(T),$$

$$c_b^2 = F_b(T).$$

F_f finds “chains” of consecutive entries of “1” and determines the length of the longest chain; F_b finds chains of consecutive “-1” and determines length of the longest chain of -1 values. Thus, c_f^2 and c_b^2 give the longest chains of monotonously ascending or descending selected template frames.

The model decides in favor of a forward movement if

$$c_f^2 > c_b^2$$

and for a backward movement if

$$c_f^2 < c_b^2.$$

For $c_f^2 = c_b^2$, the model randomly decides in favor of forward or backward movement.

This procedure is applied to each of the 100 trials of a simulation run, and the proportion of correct decisions is expressed as percentage correct.

Acknowledgments

This work was supported by the BioFuture prize of the German Federal Ministry of Education and Research. We thank Jaap A. Beintema for helpful comments on the manuscript.

Commercial relationships: none.

Corresponding author: Markus Lappe.

Email: mlappe@psy.uni-muenster.de.

Address: Department of Psychology, Institute II, Fliednerstr, 21, 48149 Münster, Germany.

References

- Beintema, J. A., Georg, K., & Lappe, M. (in press). Perception of biological from limited lifetime stimuli. *Perception & Psychophysics*.
- Beintema, J. A., & Lappe, M. (2002). Perception of biological motion without local image motion. *Proceedings of the National Academy of Sciences of the United States of America*, 99, 5661–5663. [PubMed] [Article]
- Bertenthal, I. B., & Pinto, J. (1994). Global processing of biological motions. *Psychological Science*, 5, 221–225.
- Casile, A., & Giese, M. A. (2005). Critical features for the recognition of biological motion. *Journal of Vision*, 5(4), 348–360, <http://journalofvision.org/5/4/6/>, doi:10.1167/5.4.6. [PubMed] [Article]
- Cavanagh, P. (1992). Attention-based motion perception. *Science*, 257, 1563–1565. [PubMed]
- Chatterjee, S. H., Freyd, J. J., & Shiffrar, M. (1996). Configural processing in the perception of apparent biological motion. *Journal of Experimental Psychology: Human Perception and Performance*, 22, 916–929. [PubMed]
- Coltheart, M. (1980). Iconic memory and visible persistence. *Perception & Psychophysics*, 27, 183–228. [PubMed]
- Cutting, J. E. (1978). A program to generate synthetic walkers as dynamic point-light displays. *Behavioral Research Methods and Instrumentation*, 10, 91–94.
- Cutting, J. E. (1981). Coding theory adapted to gait perception. *Journal of Experimental Psychology: Human Perception and Performance*, 7, 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, 353–356.
- Dittrich, W. H. (1993). Action categories and the perception of biological motion. *Perception*, 22, 15–22. [PubMed]
- Giese, M. A., & Poggio, T. (2000). Morphable models for the analysis and synthesis of complex motion patterns. *International Journal of Computer Vision*, 38, 59–73.
- Giese, M. A., & Poggio, T. (2003). Neural mechanisms for the recognition of biological movements. *Nature Reviews: Neuroscience*, 4, 179–192. [PubMed]
- Johansson, G. (1973). Visual perception of biological motion and a model for its analysis. *Perception & Psychophysics*, 14, 201–211.
- Lee, J., & Wong, W. (2004). A stochastic model for the detection of coherent motion. *Biological Cybernetics*, 91, 306–314. [PubMed]
- Loula, F., Prasad, S., Harber, K., & Shiffrar, M. (2005). Recognizing people from their movement. *Journal of Experimental Psychology: Human Perception and Performance*, 31, 210–220. [PubMed]
- Lu, Z. L., & Sperling, G. (1995). Attention-generated apparent motion. *Nature*, 377, 237–239. [PubMed]
- Marr, D., & Nishihara, H. K. (1978). Representation and recognition of the spatial organization of three-

- dimensional shapes. *Proceedings of the Royal Society of London: Series B*, 200, 269–294. [[PubMed](#)]
- Mather, G., Radford, K., & West, S. (1992). Low-level visual processing of biological motion. *Proceedings: Biological Sciences/The Royal Society*, 249, 149–155. [[PubMed](#)]
- Neri, P., Morrone, M. C., & Burr, D. C. (1998). Seeing biological motion. *Nature*, 395, 894–896. [[PubMed](#)]
- Pollick, F. E., Fidopiastis, C., & Braden, V. (2001). Recognising the style of spatially exaggerated tennis serves. *Perception*, 30, 323–338. [[PubMed](#)]
- 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, 2345–2355. [[PubMed](#)]
- Shiffrar, M., Lichtey, L., & Heptula Chatterjee, S. (1997). The perception of biological motion across apertures. *Perception & Psychophysics*, 59, 51–59. [[PubMed](#)]
- Sinha, P., & Poggio, T. (1996). Role of learning in three-dimensional form perception. *Nature*, 384, 460–463. [[PubMed](#)]
- Thornton, I. M., Pinto, J., & Shiffrar, M. (1998). The visual perception of human locomotion. *Cognitive Neuropsychology*, 15, 535–552.
- Thornton, I. M., Rensink, R. A., & Shiffrar, M. (2002). Active versus passive processing of biological motion. *Perception*, 31, 837–853. [[PubMed](#)]
- Thornton, I. M., & Vuong, Q. C. (2004). Incidental processing of biological motion. *Current Biology*, 14, 1084–1089. [[PubMed](#)] [[Article](#)]
- Troje, N. F. (2002). Decomposing biological motion: A framework for analysis and synthesis of human gait patterns. *Journal of Vision*, 2(5), 371–387, <http://journalofvision.org/2/5/2/>, doi:10.1167/2.5.2. [[PubMed](#)] [[Article](#)]
- Troje, N. F., & Westhoff, C. (2006). The inversion effect in biological motion perception: Evidence for a “life detector”? *Current Biology*, 16, 821–824. [[PubMed](#)]
- Troje, N. F., Westhoff, C., & Lavrov, M. (2005). Person identification from biological motion: Effects of structural and kinematic cues. *Perception & Psychophysics*, 67, 667–675. [[PubMed](#)]