Human efficiency in detecting and discriminating biological motion

Hongjing Lu
Department of Psychology, University of California, Los Angeles, Los Angeles, CA, USA
Department of Statistics, University of California, Los Angeles, Los Angeles, CA, USA

Bosco S. Tjan
Department of Psychology and Neuroscience Graduate Program, University of Southern California, Los Angeles, CA, USA

Zili Liu
Department of Psychology, University of California, Los Angeles, Los Angeles, CA, USA

Using an "information meter" provided by ideal observer analysis, we measured the efficiency with which human observers processed different walking stimuli against luminance noise and spatial uncertainty to either detect the presence of a walker or to discriminate the walking direction. Human efficiency was examined across four renderings of a human walker: contour, point lights, silhouette, and skeleton. We replicated the previous finding of low discrimination efficiency in biological motion (Gold, Tadin, Cook, & Blake, 2008) and also found low detection efficiency for biological motion. Interestingly, in *both* detection and discrimination tasks, the skeleton display was among those yielding the highest level of efficiency in processing visual information. This finding suggests that structural information about the relative position of joints, highlighted in the skeleton display, provides a critical component of the internal representation for biological motion.

Introduction

Humans show remarkable ability to recognize objects from two-dimensional retinal inputs despite drastic changes in their appearance. The ability to recognize nonrigid objects is particularly challenging because articulation and deformation entail a large space of object configurations. A prime example of a nonrigid object is the moving human body, which involves many degrees of freedom in limb movements and articulated body structure. Perception of biological motion probably underlies what may be the most sophisticated form of visual recognition processing. However, people appear to readily recognize human body movements (a special case of biological motion) from very sparse visual input, such as a dozen disconnected dot movements in a point light display (Johansson, 1973).

Indeed, the human visual system can perceive a variety of actions presented in a point light display in an automatic, effortless, and robust manner. For example, recognition of an action-in-motion sequence can be achieved from a point light display as brief as 200 ms (Johansson, 1973), when contrast is assigned randomly to each point light in each frame (Ahlstrom, Blake, & Ahlstrom, 1997), when biological motion is defined by texture (second-order motion) instead of luminance (first-order motion; Ahlstrom et al., 1997; Bellefeuille & Faubert, 1998), when only subconfigurations of a human figure are presented as point lights (Neri, 2009; Pinto & Shiffrar, 1999), when biological motion is presented across apertures (Lu, 2010; Shiffrar, Lichtey, & Heptulla Chatterjee, 1997), when biological motion is masked by random-dot kinematograms (Bertenthal & Pinto, 1994; Cutting & Kozlowski, 1977; Neri, Morrone, & Burr, 1998), and when biological motion stimuli are presented in the periphery (Thompson, Hansen, Hess, & Troje, 2007; Thurman & Lu, 2013a; van Boxtel & Lu, 2011).

The remarkably rapid, accurate, and robust perception of biological motion appears to imply that the human visual system is highly efficient in processing impoverished visual inputs provided in point light displays. However, as pointed out by Gold et al. (2008), the claim of high efficiency in human perception of biological motion cannot be warranted in the absence of a method to precisely quantify how much informa-

Citation: Lu, H., Tjan, B. S., & Lui, Z. (2017). Human efficiency in detecting and discriminating biological motion. *Journal of Vision, 17*(6):4, 1–14, doi:10.1167/17.6.4.

doi: 10.1167/17.6.4 Received December 24, 2016; published June 7, 2017 ISSN 1534-7362

tion is in fact contained in a point light stimulus. Good recognition performance, usually measured by accuracy or reaction time, may result from an intelligent system that efficiently processes visual signals despite the input being impoverished, noisy, or sparse. Alternatively, good recognition performance might arise from limited but sufficient input information that is given to an inefficient system. To disentangle these two potential explanations for the apparent efficiency of human observers in recognizing biological motion, it is necessary to measure (a) the informativeness of a visual input, and (b) the efficiency with which the human visual system processes this input information to make recognition judgment.

It is therefore essential to work within a theoretical framework that quantifies the amount of visual information. The ideal observer (IO) approach, based on Bayesian statistical inference, is a well-established method to address this problem (Geisler, 2002; Green & Swets, 1966). An IO can serve as an "information meter" to compute the optimal performance for a specific task, which then provides a quantitative measure of the stimulus information embedded in the input (e.g., Tjan, Braje, Legge, & Kersten, 1995). By comparing human performance to the ideal, one can calculate the *statistical efficiency* of the human visual system, thereby quantifying the efficiency of the human visual system in extracting, representing, and utilizing key information in the visual stimulus to perform a well-defined task.

Gold et al. (2008), for example, added dynamic luminance noise onto image frames depicting walking actions and compared the amount of information contained in a display of point lights to that of the corresponding full-figure display. They found that IO performance with point lights was about the same as with the full-body display, indicating that the two types of displays provided a similar amount of information relevant to the discrimination of walking direction. However, human efficiency in identifying biological motion was rather low in absolute terms (approximately 0.4% in efficiency for point lights vs. 2.5% for full-figure stimuli), which challenged the common dogma that humans are efficient in processing biological motion (given their high accuracy in recognizing actions from sparse stimuli such as the point light display).

The present work extends the study by Gold et al. (2008) to examine efficiencies of visual processing across different input formats of biological motion and different tasks under spatial uncertainty. In previous work, four types of renderings for action stimuli have been used to study biological motion perception. The same human body movements can be displayed using different renderings: contour, point light, silhouette, and skeleton. Each of the rendering stimuli provides different types of features in the visual input. For example, the full-figure condition in Gold et al. (2008), a silhouette, provides information about body shape and associated movements. As the most commonly used display in studies of biological motion perception, point light stimuli appear to provide the sparsest and the most compact information about the kinematics involved in human body movements by only showing joint movements. Contour-based displays, which contain information about boundaries, have been widely used in computer vision to track and recognize human movements from raw video frames (Blake & Isard, 1998). However, it is unclear whether the human visual system processes contour information efficiently. The skeleton stimulus has been used less commonly in psychophysical investigations of biological motion as most studies have emphasized point light displays. However, a skeleton conveys critical structural information about the body, including not only joint positions but also the *relationships* between joints shown as connections (Feldman & Singh, 2006). At the same time, the skeleton display, like the point light display, eliminates detailed body shape. In previous research, human ability to recognize actions regardless of the rendering conditions of the stimuli has been taken as evidence of the robustness of biological motion perception. However, whether humans are equally efficient in processing the same action rendered in different ways remains unclear. If, on one hand, biological motion perception depends on some visual features that are salient in one rendering stimulus but less prominent in others, we would expect that human efficiency in processing the visual information may vary across different rendering conditions. On the other hand, if biological motion perception depends on a more abstract representation of actions that is independent of visual features, we would expect the same efficiency in perception of biological motion across the different rendering conditions. Thus by systematically examining the distinct stimuli based on different rendering conditions, we may be able to shed light on the internal representation of biological motion that the visual system uses and determine the certain critical features on which the system operates.

In addition to varying the rendering of the action inputs, we also manipulated the task (Abbey & Eckstein, 2006) that observers were required to perform under spatial uncertainty for biological motion perception. In studies on biological motion perception, two tasks have been used: *detection* tasks in which observers judge whether the input contains an exemplar of a visual type (e.g., a human walker) or contains pure dynamic noise and *discrimination* tasks that require the observer to make a more fine-grained discrimination between subtypes of different actions (e.g., deciding whether the walker was moving left or right). The two

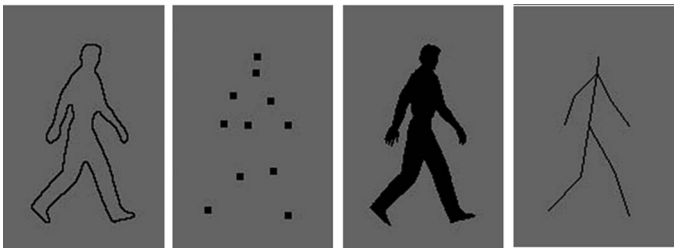

Figure 1. Illustration of four stimulus renderings: contour, point light, silhouette, and skeleton walker, each in a static frame.

tasks are related in some ways. When the display position of the target walker is unpredictable, participants may need to first detect where the walker is shown in the display and then recognize the walking direction. If the processes take place in such a serial manner, the visual processes underlying detection of biological motion may constitute the initial processes required for recognition of biological motion, in which case detection efficiency would impose an upper limit on the efficiency of recognizing biological motion. However, previous research has shown that detection and discrimination tasks can probe different types of visual features in the stimuli (Neri & Heeger, 2002). If discrimination of walking direction operates on a different set of features, which are not used for detection, it would be possible for discrimination efficiency to exceed detection efficiency in some conditions. The present study employed both detection and discrimination tasks in order to explore these possibilities.

We report two experiments that measured human efficiencies with four displays of biological motion in a detection task and a discrimination task, respectively. The findings from the present study will allow us to assess how efficiently humans detect/discriminate biological motion compared to other stimuli, such as simple shapes, letters, and rigid objects. These results will also shed light on the critical information used in forming location-invariant representations of biological motion to enable robust performance across different tasks.

General methods

Apparatus

Stimuli were presented on a Dell monitor with a refresh rate of 75 Hz and resolution of 1024 × 768. At the viewing distance of 57 cm (maintained via a chin rest), each pixel subtended 1.62 arcmin. The monitor was calibrated with a Minolta CS-100 photometer. A lookup table was constructed to allow linear division of a luminance range, 1.96~170 cd/m^2, into 256 programmable intensity levels. Experiments were conducted in a dim room. We used Poser 4 software (MetaCreation Inc.) and Photoshop to create the stimuli and Matlab (MathWorks Inc.) and PsychToolbox (Brainard, 1997; Pelli, 1997) to present the stimuli.

Stimuli

Each stimulus display contained a dark target, namely a human walker, on a gray (46.50 cd/m^2) window. The display window was centered on a black (1.96 cd/m^2) screen. The size of the display window was 120 pixels horizontally by 180 pixels vertically, subtending 3.24° by 4.96° of visual angle. The target was walking as if on a treadmill in one walking cycle (including two steps), which consisted of 10 image frames presented at a rate of 133 ms/frame. The center location of the target in each trial was randomly selected in a window of 20 × 30 pixels in size (0.54° × 0.81°). The target in both experiments depicted a human walker in one of the four rendering conditions: contour, point light, silhouette, and skeleton in the absence of noise (see Figure 1). All the four stimulus types used two gray levels, one for the foreground and another for the background. The luminance of the background was constant, 46.50 cd/m^2, and the luminance of the foreground varied depending on the contrast in each trial as determined by a staircase procedure.

A silhouette walker was saved frame by frame using MetaCreation Poser software under orthographic projection. All pixels inside the bounding contour were set to a uniform dark gray (foreground). The walker subtended about 90 pixels horizontally by 140 pixels vertically (2.43° × 3.86°). The average number of foreground pixels per frame was 3,086 in the silhouette stimulus, corresponding to 14.3% pixels in the display window.

A contour stimulus was produced by using Photoshop to mark all of the luminance discontinuities in a silhouette stimulus by lines with width of one pixel. As was the case for the silhouettes, a darker gray for the bounding contour (foreground) and a brighter gray for the background were used to generate the contour stimulus and the other two types of stimuli. The average number of foreground pixels per frame in a contour stimulus was 566, corresponding to 2.6% pixels in the display window.

A point light stimulus was generated based on orthogonal projections of the known three-dimensional coordinates of 13 joints: head, left/right shoulder, left/right elbow, left/right hand, left/right thigh, left/right knee, and left/right foot. Each point light (foreground)

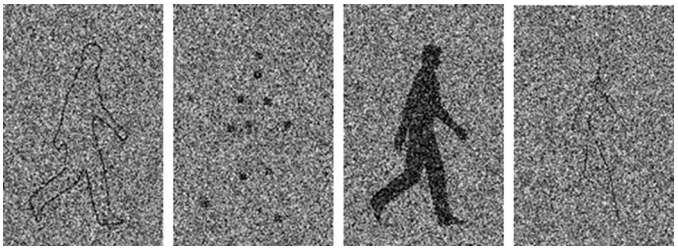

Figure 2. Illustration of a static frame in each of the noisy stimulus renderings: contour, point light, silhouette, and skeleton.

was displayed as a square with the size of 5×5 pixels ($0.14° \times 0.14°$). A total of 13 point lights were displayed in dark gray against a background of brighter gray. The average number of foreground pixels per frame was 271, corresponding to 1.3% pixels in the display window.

A skeleton stimulus was produced by connecting two joints in accord with the human body hierarchy. The line width of the skeleton stimulus was the same as in the contour stimulus, one pixel wide. The average number of foreground pixels per frame was 258, corresponding to 1.2% pixels in the display window.

Dynamic Gaussian luminance noise fields were generated independently in spatial and temporal dimensions (see Figure 2). Noise fields sampled from a Gaussian distribution of zero intensity and σ standard deviation were superimposed on each pixel in target frames. Two noise levels were used with *SD* of 3.59 and 12.10 cd/m^2 for practice and experimental tests, respectively. The two noise levels resulted in noise spectral density of 1.04 and 16.60 μ(deg^2 sec) or 10^{-6} deg^2 sec, respectively.

Procedure

Prior to testing, participants were presented with four types of walking stimuli in the absence of luminance noise for six walking cycles. Participants were informed that these four types of stimuli were generated from the same movement of a person. Participants then received a practice session with the target walker in the absence of luminance noise. The four types of stimuli were presented at three frame rates: 133 ms/frame, 266 ms/frame, and 399 ms/frame. The practice session consisted of 24 trials (four rendering conditions, three frame rates, presence/absence of the walker in Experiment 1 or two walking directions in Experiment 2 with feedback based on the appropriate judgment (detection in Experiment 1, discrimination in Experiment 2).

One experimental run consisted of four sessions, each including one rendering condition. The session sequence was counterbalanced between participants. Each session consisted of two blocks: 12 trials in a practice block with feedback and 300 trials in the experimental block without feedback. The trials in the practice block were easy with low noise background (spectral density 1.04 μ(deg^2 sec)) and three levels of target luminance (0.00, 7.55, and 15.10 cd/m^2, four trials each), showing high-contrast stimuli. If a participant made more than a single error, an additional 12 practice trials were added. Between trials, participants were presented a uniform gray display of 46.50 cd/m^2, equal to the mean luminance of the noisy background. In the subsequent experimental block with the same rendering type, noise fields had high spectral density [16.60 μ(deg^2 sec)]. An adaptive psychophysical staircase procedure, QUEST (Watson & Pelli, 1983), was used to adjust contrasts of the target walker to yield 75% accuracy. The next trial started 1 s after the participant made a response by pressing a key.

Experiment 1: Detection of biological motion

Method

Rightward-walking targets embedded in luminance noise were displayed on half of the trials, and only luminance noise fields were presented in the other half. Before the experiment started, participants were informed that the walking direction of targets was always toward the right. Participants were asked to detect the human walker and to respond by pressing one button indicating "present" and another indicating "absent."

Six undergraduate students at the University of California, Los Angeles (UCLA) participated in Experiment 1 for course credit and in accordance with the Helsinki Declaration.

IO simulation

The IO analysis is presented in the Appendix. The rightward-walking targets were known to the IO, but the display location on each trial was unknown. The IO consequently considered all possible locations, summing over them by treating the target location as a hidden variable. Spatial uncertainty only affected the formulation of the likelihood when a target was present, not the decision rule nor the likelihood when a target was absent (see Equations 3 and 4 in the Appendix).

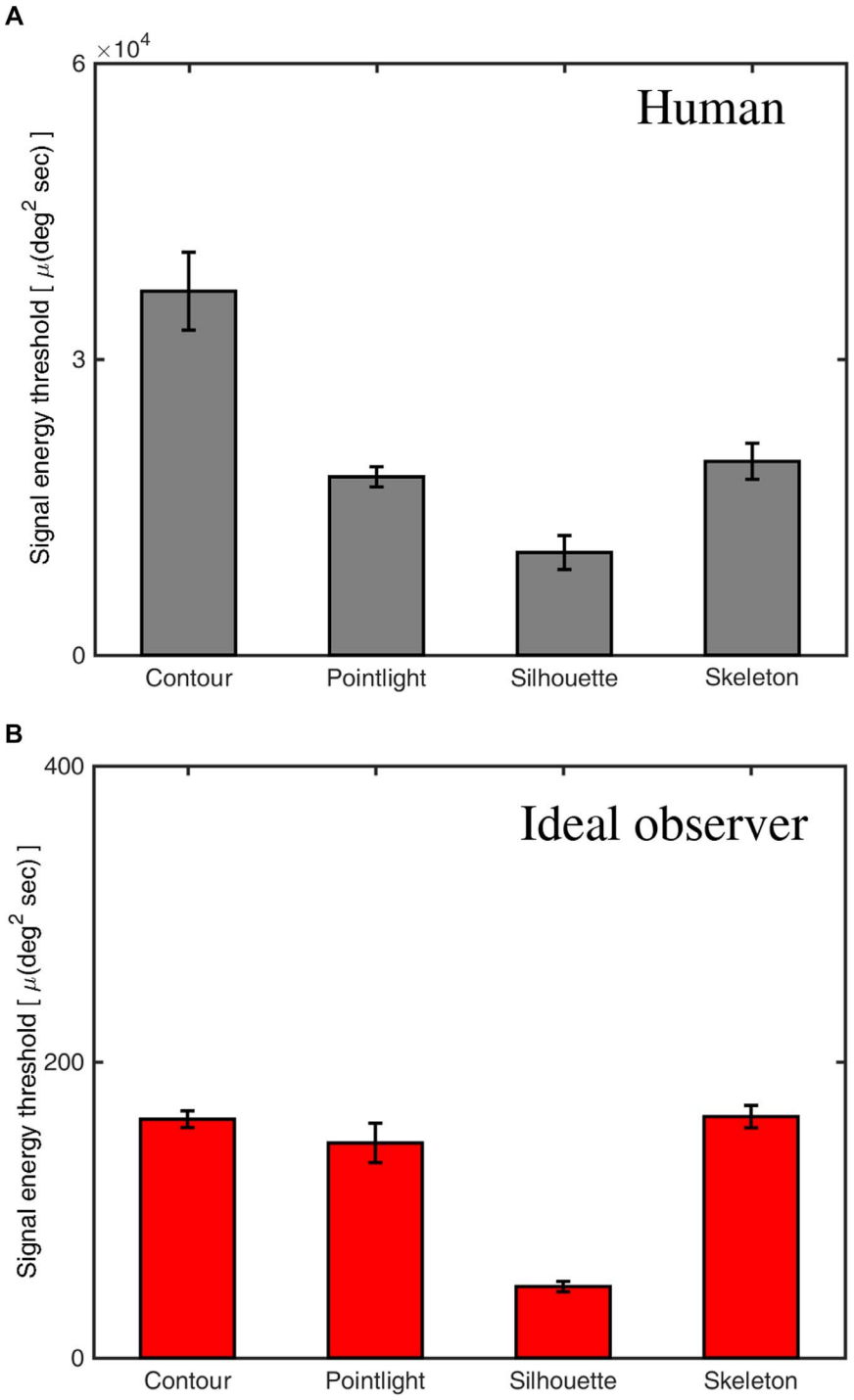

Figure 3. Contrast thresholds at 75% correct for (A) human observers and (B) the IO in the detection task (Experiment 1). Error bars indicate one *SEM*.

Results for human observers and the IO

Contrast thresholds at 75% correct for human observers were compared across four types of rendering conditions. Lower contrast thresholds of targets indicate better detection performance. Human results, depicted in Figure 3A, indicated that the best detection performance was obtained in the silhouette condition ($M = 0.010$ deg^2 sec, $SD = 0.0042$), followed by point light ($M = 0.018$, $SD = 0.0025$) and skeleton ($M = 0.020$, $SD = 0.0045$) conditions with the worst performance in the contour condition ($M = 0.036$, $SD = 0.0097$). All pairwise comparisons were significant after Bonferroni adjustment for multiple comparisons ($p < 0.05$) except the contrast between point lights and skeleton.

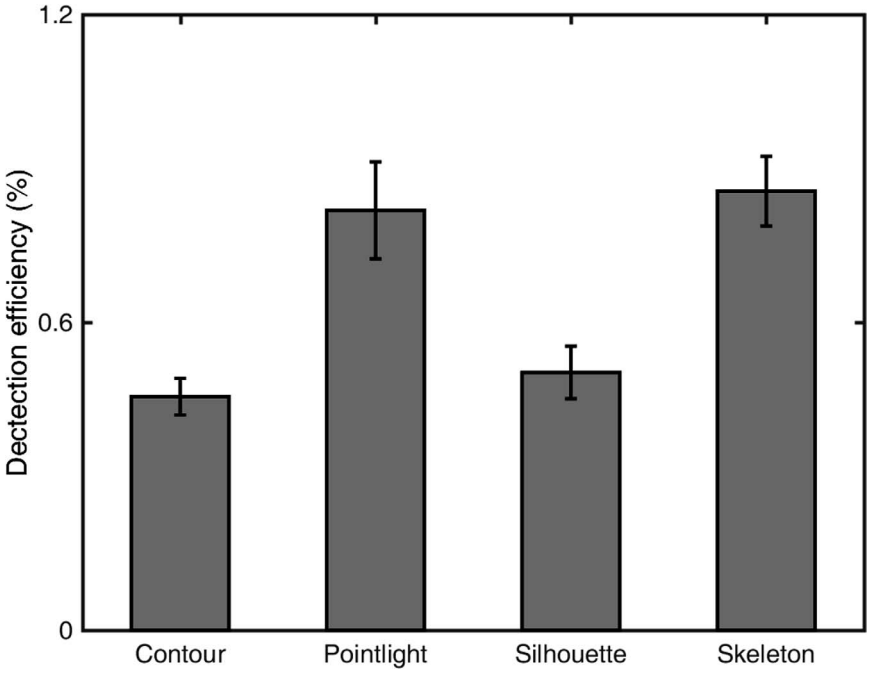

Figure 4. Human detection efficiencies in the detection task (Experiment 1) with four rendering conditions: contour, point light, silhouette, and skeleton.

Figure 3B shows IO performance for each condition ($M = 161.19$, 145.18, 48.13, 162.97 μdeg^2 sec for contour, point light, silhouette, and skeleton, respectively). Based on this "information meter," the IO simulation clearly showed that silhouettes (which yielded the lowest ideal contrast threshold) provided more information for detecting the walker than did the other three conditions.

As a quantitative measure of how efficiently the human visual system used the input information in the four rendering conditions, we computed *detection efficiency*: the ratio of signal energy threshold for the IO and human observers to perform the same detection task. As shown in Figure 4, efficiency of detecting the target walker was higher for point light ($M = 0.82\% \pm SD = 0.23\%$) and skeleton displays ($0.86\% \pm 0.17\%$) than for contour ($0.46\% \pm 0.09\%$) and silhouette displays ($0.50\% \pm 0.13\%$). A Bonferroni multiple comparison test revealed significant differences in human detection efficiency for contour versus point lights, contour versus skeleton, point lights versus silhouette, and point lights versus skeleton, $t(5) > 2.5$, $p < 0.02$, $d > 1.6$ for all comparisons. No significant difference was obtained between point light and skeleton displays, $t(5) = 0.29$, $p = 0.78$, $d = 0.19$, or between contour and silhouette displays, $t(5) = 0.82$, $p = 0.45$, $d = 0.43$.

The efficiency measures thus revealed a very important result. Although the skeleton contained the least information of the four stimulus types (based on the mean threshold of the IO) and therefore may not afford the greatest overall performance, the skeleton display nonetheless was processed most efficiently by the human visual system as shown in Figure 4. A similar result was also found for the point light display in the detection task.

In comparison, the efficiency for the silhouette was not very high. It seems that although the IO used every available signal pixel in the stimulus for detecting actions, human participants could not use all the signal pixels within the silhouette. In other words, there might be too many signal pixels for humans to effectively use.

A very interesting result is that the point lights had an efficiency nearly as high as the skeleton. Given that the statistical efficiency had already taken into consideration the objectively available stimulus information, this relatively high efficiency could reflect the information encoded in the human internal representation for detection. This result could not have been obtained without comparing across these display types.

Discussion

It may be helpful to provide intuitive explanations regarding the IO performance shown in Figure 3B. Recall that because the IO did not know the precise location of the walker in each trial, it had to consider all possible positions and integrate them to calculate the likelihood. It did know that the walker would be always upright. The IO's template matching from one position to the next was a translational motion. In light

of this fact, it is not surprising that the silhouette gave rise to the lowest threshold. In other words, it was the easiest to detect. This is because, when the IO's template moved away from perfectly matching the walker's position, there was still partial overlap between the template and the silhouette walker. Therefore, these large overlaps still contributed meaningfully to increase the likelihood in the presence of the target walker. The larger the object area is, as in the case of the silhouette, the greater the overall likelihood will be because there are many partial matches.

Along the same lines, we can also qualitatively reason why the point lights and skeleton showed comparable thresholds for the IO. In the case of the point lights, because each point had the size of 5×5 pixels, there was still partial overlap between the template and the stimulus so long as the template translated within this window of 5×5 pixels around the position of the perfect match. Likewise, in the case of the skeleton, because the lines connecting the joints were straight, these lines also entailed partial translational invariance. In other words, when the template moved along the orientation of one of the lines from the perfect matching position, there would be still partial overlap. In comparison with the silhouette, the point lights and skeleton stimuli yielded higher thresholds (i.e., poorer performance) because the overlap areas for the point lights and skeleton were smaller than those for the silhouette.

Once these three cases are understood qualitatively, it is easier to understand the case of the contour. To reiterate, the likelihood term depends on two factors: the target area and configuration. If the configuration is partially translation-invariant, an offset from the perfect match still contributes meaningfully to the likelihood calculation. Because the contour was mostly curved and one pixel in width, it had little translational invariance. Accordingly, when the template and the stimulus did not perfectly match, the overlap would be small. That is why the contour's threshold for the IO was comparable with those of the point light and skeleton even though the contour's area was about twice as large.

The qualitative description above is meant to help understand why the IO's thresholds in the four renderings depend on both the area and the configuration of an object. These four thresholds shown in Figure 3B were obtained from the IO computation and not human behavior. The important point to emphasize is that by the nature of an object's area and configuration, by the nature of the noise generation, and by the spatial uncertainty defined in this task (in which the IO must consider all possible positions but not orientations of an object), different stimulus renderings encode different amounts of visual signal and thus give rise to different detection thresholds. Understanding the IO thresholds should help interpret human detection thresholds, and the corresponding statistical efficiencies.

Experiment 2: Discrimination of biological motion

Method

Two side views (0° and 180°) were used to generate leftward-walking and rightward-walking sequences. On each trial, either leftward or rightward walking was randomly selected with equal probability. The target sequence was embedded in dynamic luminance noise as in Experiment 1. The participants were asked to discriminate the walking direction of the nonrigid motion pattern of the walker. The design and procedure were otherwise the same as in Experiment 1.

Participants were asked to identify the target as "leftward walking" or "rightward walking" by pressing one of two buttons. We recruited six UCLA undergraduate students who had not participated in the other experiment. They were given course credit for their participation in the study.

IO simulation

The IO simulation (see Appendix) was based on Equations 5 and 6, which sum over all possible spatial locations of the target as a hidden variable. Spatial uncertainty affects the formulation of likelihoods for both the left- and the right-walking targets. The IO used the decision rule specified in Equation 7 to make a discrimination response.

Results for human observers and the IO

As depicted in Figure 5A, the measure of signal contrast threshold for human participants revealed the best discrimination performance in the silhouette condition ($M = 0.02 \pm SD = 0.008$), the worst performance in the contour (0.048 ± 0.006) and point light conditions (0.045 ± 0.016), and intermediate performance in the skeleton condition (0.027 ± 0.006). All pairwise comparisons were significant after Bonferroni adjustment for multiple comparisons ($p < 0.04$) except the comparisons between contour versus point lights and silhouette versus skeleton.

For comparison, Figure 5B displays the corresponding IO performance for each condition, providing a measure of the amount of information embedded in

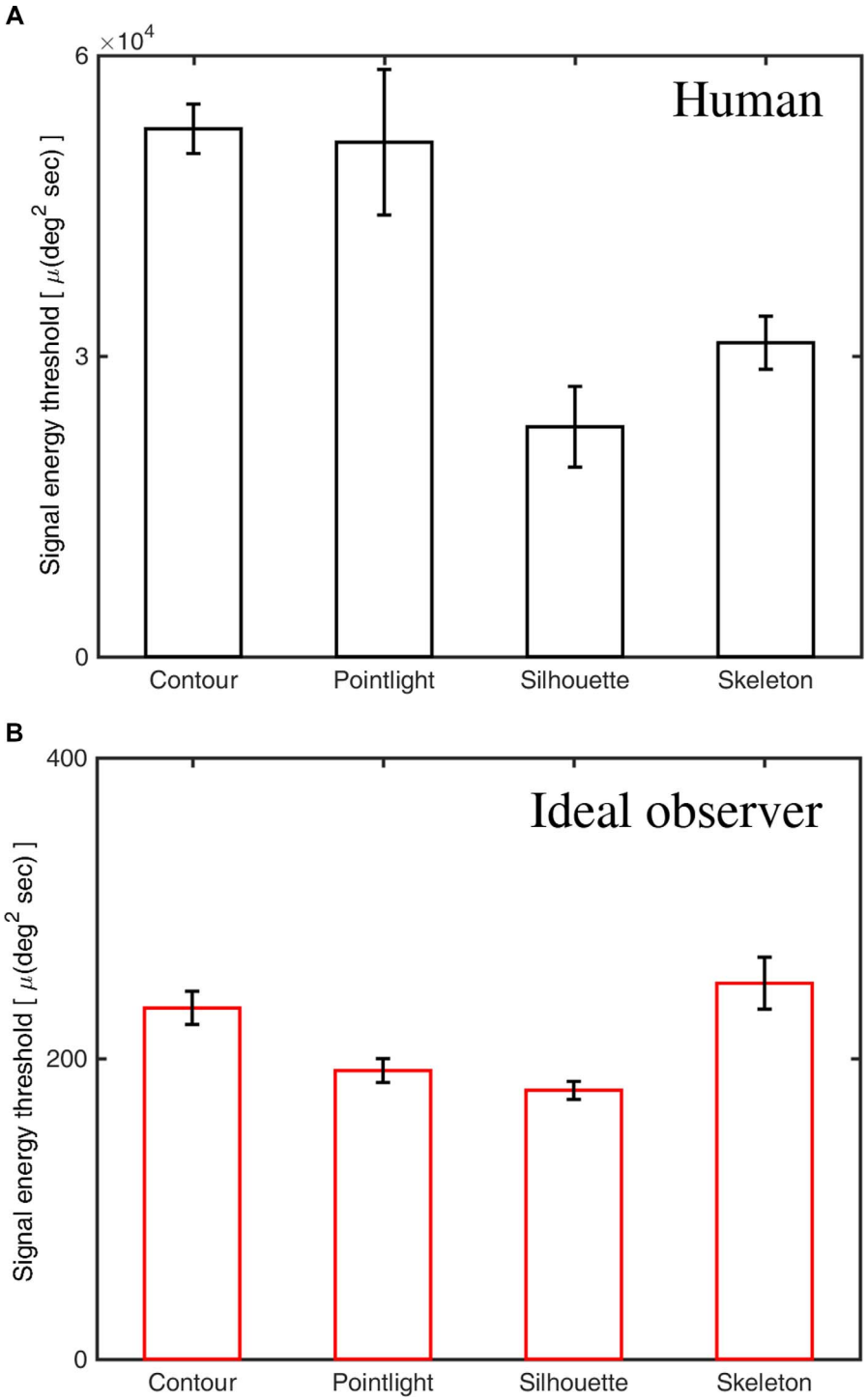

Figure 5. Contrast thresholds at 75% correct for (A) human observers and (B) the IO in the discrimination task (Experiment 2). Error bars indicate one *SEM*.

each display for the task of walking direction discrimination. The IO performance showed that the greatest amount of information was provided in the silhouette display as revealed by the lowest signal energy threshold whereas the least information was provided in the skeleton display. One intuitive explanation is that silhouette display provided a large number of non-overlapping pixels between leftward and rightward walkers so that the visual cues entailed more discriminative information between the two templates. On the other hand, the skeleton display may share proportionally more pixels between the two walkers with the opposite-facing directions (e.g., the head pixels, the torso pixels were very close between the two templates).

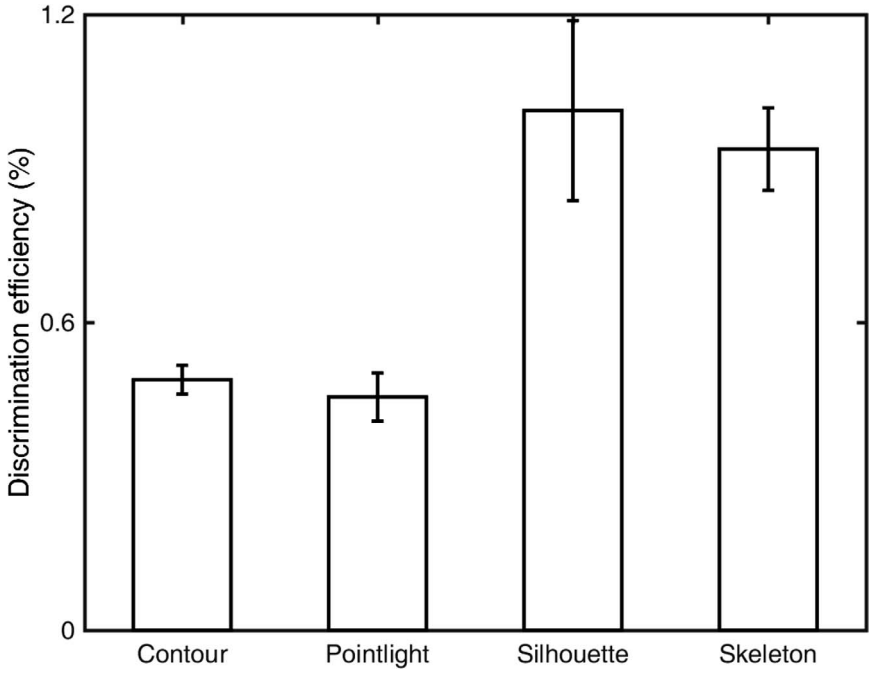

Figure 6. Human efficiencies in the discrimination task (Experiment 2) with four rendering conditions: contour, point light, silhouette, and skeleton.

Hence more ambiguous information was included in the skeleton display, which resulted in worse performance.

Note that the ideal thresholds for point lights and silhouette were similar, consistent with the comparable finding in a discrimination task reported by Gold et al. (2008). However, we found a slightly higher threshold for the point lights than for the silhouette whereas Gold et al. found the same threshold for the two displays. This small discrepancy was likely due to the difference in the number of frames in the walking sequence (we used 10 frames whereas Gold et al. used 20 frames). In addition, the present study used a larger human figure and a smaller location shift window.

Figure 6 depicts the efficiencies by comparing human discrimination threshold with ideal performance. The efficiency measures revealed that humans achieved highest discrimination efficiency for silhouette ($M = 1.01\% \pm 0.42\%$) and skeleton ($0.94\% \pm 0.19\%$). There was no significant difference between these two conditions, $t(5) = 0.32$, $p = 0.77$, $d = 0.22$. A Bonferroni multiple comparison test further confirmed that human discrimination efficiency for skeleton was higher than for contour and point lights, $t(5) > 2$, $p < 0.03$, $d > 1.7$ for all comparisons. This result revealed that the human performance difference between silhouettes and skeletons (see Figure 5A) was attributable to the greater amount of information contained in the silhouettes whereas processing efficiency within the human visual system was comparable for these two displays.

To combine human efficiency results from both detection and discrimination tasks across the two experiments, Figure 7 depicts the results of a cross-task comparison. The silhouettes yielded discrimination efficiency (1%) significantly higher than detection efficiency (0.5%), $t(10) = 2.84$, $p = 0.02$, $d = 1.64$. Discrimination efficiency was not significantly higher than detection efficiency for contours (0.49% vs. 0.46%), $t(10) = 0.72$, $p = 0.49$, $d = 0.42$, and skeletons (0.94% vs. 0.86%), $t(10) = 0.76$, $p = 0.46$, $d = 0.44$. Only for point lights, discrimination efficiency (0.45%) was significantly lower than detection efficiency (0.82%), $t(10) = 3.48$, $p < 0.01$, $d = 2.01$. These results imply that detection efficiency does not necessarily impose an upper limit on discrimination efficiency under conditions of spatial uncertainty across different rendering conditions. This finding suggests that detection and discrimination may use different types of features in processing the visual information in some rendering conditions (e.g., silhouettes).

General discussion

Using the "information meter" provided by the IO analysis, we measured the efficiency with which human observers processed each stimulus type against luminance noise to either detect the presence of a walker (Experiment 1) or discriminate the walking direction of the walker (Experiment 2). In both tasks, the skeleton

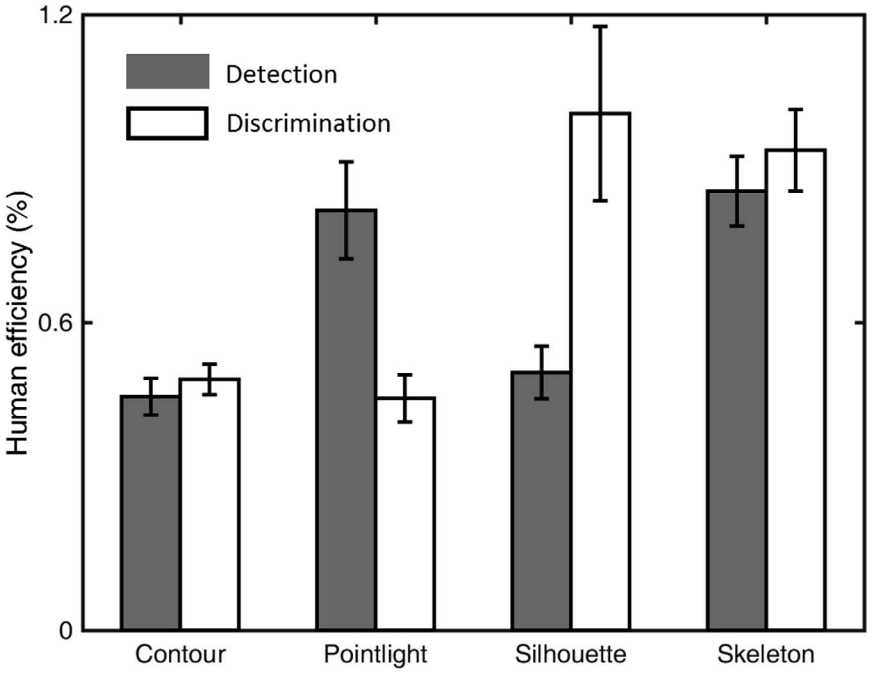

Figure 7. A comparison of human detection efficiency (gray bars, Experiment 1) and discrimination efficiency (white bars, Experiment 2) with the four rendering conditions.

display was among those that yielded the highest processing efficiency. This finding suggests that structural information about the relative relationships of joints, highlighted in the skeleton display, provided a critical component of the internal representation for biological motion. Moreover, although both contour and skeleton displays involved line drawings, discrimination efficiency was lower for contours than for skeletons, indicating that humans were able to use structural information more efficiently when it was explicitly presented as joint connections in a skeleton display. In addition, although both contour and silhouette displays included information about the shape of the human body, the lower discrimination efficiency observed for contours than for silhouettes suggests that humans may not efficiently use boundary information about body shape in recognizing biological motion.

The skeleton display was the only stimulus that yielded the superior efficiency among the four display types in *both* detection and discrimination tasks. Note that the silhouette displays also yielded the highest efficiency in the discrimination task but not in the detection task. This superior efficiency for processing skeleton stimuli across two different tasks suggests that the structural information about relative limb connections provided by the skeleton displays was especially important for the internal representation that humans used for biological motion perception. The importance of structural processing in biological motion perception has been supported by previous studies using different experimental manipulations, such as body orientation (Sumi, 1984), body subparts (Pinto & Shiffrar, 1999), stimuli shown across apertures (Lu, 2010), and scrambled biological motion stimuli (Thurman & Lu, 2013b). Neuroimaging studies have also provided evidence to support complementary but dissociable neural mechanisms underlying structural processing and motion processing in perception of bodily movements (Vangeneugden, Peelen, Tadin, & Battelli, 2014). The IO analysis in the present study provides further converging evidence, highlighting the important role of efficient structural processing or form processing in biological motion detection and recognition.

We also found that, when the location of the target display varied randomly on each trial, discrimination efficiency for the silhouette was significantly greater than detection efficiency. This finding indicates that location-invariant recognition of biological motion for some types of renderings may involve different types of features than those used in detection. However, efficiency in discriminating point light stimuli remained lower than detection efficiency, which was consistent with the hypothesis that, for this type of rendering, discrimination may indeed be constrained by efficiency of detecting local features, such as individual point lights.

Was the pattern of results across the rendering conditions dependent critically on the spatial jitter used in the experiments? To address this question, we ran additional simulations to examine the impact of the jitter range on the IO performance using a range of

jitter sizes: 6×4, 12×8, 18×12, 24×16, and 30×20 pixels (which was the largest possible jitter range in the current setup for the display window). We found that the relative IO performance across the four rendering conditions remained similar across all five jitter ranges. For example, in the detection task, the lowest threshold obtained was for the silhouette condition, followed by the point light. The highest thresholds were obtained for the contour and skeleton conditions. In the discrimination task, the point light and silhouette conditions consistently showed lower thresholds than the contour and skeleton conditions across the tested jitter range. Given that similar trends were maintained across a range of spatial jitters, the results regarding relative efficiency found in the present study were likely generalizable to other spatial jitters.

A pervasive finding that may seem surprising, given the common assumption that the human visual system readily perceives point light displays (Johansson, 1973) is that in absolute terms human efficiencies in detecting and discriminating biological motion were quite low (less than 2%). The present results confirmed the similar low efficiency level reported by Gold et al. (2008) for discrimination of walking direction with point light and silhouette and extended the generality of their finding to two additional types of renderings and to detection as well as discrimination. It should be noted that in another IO analysis of gender recognition from human walking actions, Pollick, Kay, Heim, and Stringer (2005) obtained much higher efficiency estimates (26%–48%) on the basis of a single critical cue (shoulder width relative to hip width). The apparent discrepancy between the efficiency estimates obtained by Pollick et al. (2005) versus Gold et al. (2008) is attributable to their use of different tasks, different types of noise, and, most importantly, different IO models. In particular, Pollick et al. (2005) used *constrained* IO (Geisler, 2002; Liu, Knill, & Kersten, 1995), in which other potentially informative cues (e.g., lateral body sway) were ignored in the model. In contrast, the study by Gold et al. (2008) and the present study used an unconstrained IO, considering all stimulus information in the visual inputs.

IO analysis has been widely employed in studies of low-level vision, such as contrast discrimination (Kersten, 1987; Legge, Kersten, & Burgess, 1987), detection of dot density (Barlow, 1978), detection of mirror symmetry in random dots (Barlow & Reeves, 1979), and discrimination of coherent motion (Barlow & Tripathy, 1997; Lu & Yuille, 2006), and tasks involving high-level vision, including object recognition (Liu, Kersten, & Knill, 1999; Liu et al., 1995; Tjan et al., 1995), word recognition (Pelli, Farell, & Moore, 2003), face recognition (Gold, Bennett, & Sekuler, 1999), and action recognition (Gold et al., 2008; Pollick, Lestou, Ryu, & Cho, 2002). This rich body of work makes it possible to compare human efficiency for higher-level visual tasks with human efficiency for lower-level tasks observed in previous studies, such as recognition of a simple shape (e.g., a circular disk; Legge et al., 1987), recognition of letters and words (Burns & Pelli, 1992), and recognition of rigid objects (Tjan et al., 1995). The present study found low efficiencies for processing biological motion (less than 2%), which contrasts with the higher values reported for tasks involving simpler static stimuli. For example, efficiency has been found to be 3%~8% for recognizing rigid objects under spatial uncertainty (Tjan et al., 1995), 1%~10% for recognizing words with 2~10 letters (Pelli et al., 2003), 12%~20% for recognizing unfiltered letters (Burns & Pelli, 1992), 14% for contrast discrimination of small disks (Legge et al., 1987), and 42% for recognizing spatially filtered letters (Parish & Sperling, 1991).

These variations in efficiency levels across stimulus types follow the general trend that efficiency declines as the level of visual processing increases (e.g., from recognizing spatially filtered letters to discrimination of biological motion). This trend may seem counterintuitive, given that the human visual system appears to be adept and robust at recognizing complex objects and analyzing visual scenes. However, the trend may in fact be indicative of a general trade-off between efficiency of processing highly specific visual inputs versus adaptively processing diverse proximal stimuli that correspond to important classes of distal stimuli in the environment. For example, in the real world, humans need to achieve invariant object recognition under varied viewing conditions (e.g., changes in lighting, viewpoint, location, and articulation). Similarly, humans need to recognize actions, such as walking, across multiple individuals who vary in body shape as well as across variations in kinematics, gait, and viewing angle. In the present experiments, by contrast, a single human walker was employed as the stimulus, and efficiency was computed on the basis of the IO that had available a specific template of this individual walker.

One intriguing comparison of efficiency across various types of stimuli is that recognition efficiency for biological motion appears to be similar to efficiency of recognizing words consisting of 10~16 letters (Pelli et al., 2003): Both are less than 1%. It is possible that recognizing biological motion based on 11 joints yields a degree of relational processing comparable to the number of perceptual relationships involved in recognizing words with 10~16 letters. One direction for future work is to use statistical efficiency measures to investigate relational representations of objects by focusing on relationships between features rather than on individual features or holistic templates.

More broadly, the results of the present study encourage further applications of formal Bayesian analyses of information content in stimuli and of

human processing efficiency to complex dynamic displays depicting human action. Future work should aim to apply similar methods to understand how humans perceive less familiar patterns of biological motion, such as those involving dance and gymnastics, and motions that vary along dimensions, such as object size and temporal uncertainty, and social interactions.

Keywords: biological motion, human efficiency, ideal observer, detection/discrimination

Acknowledgments

We dedicate this paper to the memory of our co-author Dr. Bosco S. Tjan, whose brilliance, insight, and generosity made this project possible. This work was supported by NSF BCS-1353391 to HL and BCS-0617628 to ZL.

Commercial relationships: none.
Corresponding author: Hongjing Lu.
Email: hongjing@ucla.edu.
Address: Department of Psychology, University of California, Los Angeles, Los Angeles, CA, USA.

References

Abbey, C. K., & Eckstein, M. P. (2006). Classification images for detection, contrast discrimination, and identification tasks with a common ideal observer. *Journal of Vision*, *6*(4):4, 335–355, doi:10.1167/6.4.4. [PubMed] [Article]

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

Barlow, H. (1978). The efficiency of detecting changes of density in random dot patterns. *Vision Research*, *18*(6), 637–650.

Barlow, H., & Reeves, B. C. (1979). The versatility and absolute efficiency of detecting mirror symmetry in random dot displays. *Vision Research*, *19*(7), 783–793.

Barlow, H., & Tripathy, S. P. (1997). Correspondence noise and signal pooling in the detection of coherent visual motion. *Journal of Neuroscience*, *17*(20), 7954–7966.

Bellefeuille, A., & Faubert, J. (1998). Independence of contour and biological-motion cues for motion-defined animal shapes. *Perception*, *27*(2), 225–235.

Bertenthal, B. I., & Pinto, J. (1994). Global processing of biological motions. *Psychological Science*, *5*(4), 221–225.

Blake, A., & Isard, M. (1998). *Active contours: The application of techniques from graphics, vision, control theory and statistics to visual tracking of shapes in motion*. London: Springer-Verlag.

Brainard, D. H. (1997). The Psychophysics Toolbox. *Spatial Vision*, *10*(4), 433–436.

Burns, C. W., & Pelli, D. G. (1992). Recognition of letters and words in noise. *Ophthalmic and Physiological Optics*, *12*(1), 84–85.

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

Feldman, J., & Singh, M. (2006). Bayesian estimation of the shape skeleton. *Proceedings of the National Academy of Sciences, USA*, *103*(47), 18014–18019, doi:10.1073/pnas.0608811103.

Fisher, R. A. (1925, July). Theory of statistical estimation. *Mathematical Proceedings of the Cambridge Philosophical Society*, *22*(5), 700–725.

Geisler, W. S. (2002). Ideal observer analysis. In L. Chalupa & J. Werner (Eds.), *The visual neurosciences* (pp. 825–837). Boston: MIT Press.

Gold, J., Bennett, P. J., & Sekuler, A. B. (1999). Identification of band-pass filtered letters and faces by human and ideal observers. *Vision Research*, *39*(21), 3537–3560.

Gold, J., Tadin, D., Cook, S. C., & Blake, R. (2008). The efficiency of biological motion perception. *Perception & Psychophysics*, *70*(1), 88–95.

Green, D. M., & Swets, J. A. (1966). *Signal detection theory and psychophysics*. New York: Wiley.

Johansson, G. (1973). Visual perception of biological motion and a model for its analysis. *Perception & Psychophysics*, *14*, 210–211.

Kersten, D. (1987). Statistical efficiency for the detection of visual noise. *Vision Research*, *27*(6), 1029–1040.

Legge, G. E., Kersten, D., & Burgess, A. E. (1987). Contrast discrimination in noise. *Journal of the Optical Society of America A*, *4*(2), 391–404.

Liu, Z., Kersten, D., & Knill, D. C. (1999). Dissociating stimulus information from internal representation—A case study in object recognition. *Vision Research*, *39*(3), 603–612.

Liu, Z., Knill, D. C., & Kersten, D. (1995). Object classification for human and ideal observers. *Vision Research*, *35*(4), 549–568.

Lu, H. (2010). Structural processing in biological

motion perception. *Journal of Vision*, *10*(12):13, 1–13, doi:10.1167/10.12.13. [PubMed] [Article]

Lu, H., & Yuille, A. L. (2006). Ideal observers for detecting motion: Correspondence noise. *Advances in Neural Information Processing Systems*, *18*, 827–834.

Neri, P. (2009). Wholes and subparts in visual processing of human agency. *Proceedings of the Royal Society of London B: Biological Sciences*, *276*(1658), 861–869.

Neri, P., & Heeger, D. J. (2002). Spatiotemporal mechanisms for detecting and identifying image features in human vision. *Nature Neuroscience*, *5*(8), 812–816.

Neri, P., Morrone, M. C., & Burr, D. C. (1998, Oct 29). Seeing biological motion. *Nature*, *395*(6705), 894–896.

Parish, D. H., & Sperling, G. (1991). Object spatial frequencies, retinal spatial frequencies, noise, and the efficiency of letter discrimination. *Vision Research*, *31*(7), 1399–1415.

Pelli, D. G. (1997). The VideoToolbox software for visual psychophysics: Transforming numbers into movies. *Spatial Vision*, *10*(4), 437–442.

Pelli, D. G., Farell, B., & Moore, D. C. (2003, June 12). The remarkable inefficiency of word recognition. *Nature*, *423*(6941), 752–756.

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

Pollick, F. E., Kay, J. W., Heim, K., & Stringer, R. (2005). Gender recognition from pointlight walkers. *Journal of Experimental Psychology: Human Perception and Performance*, *31*(6), 1247–1265, doi: 10.1037/0096-1523.31.6.1247.

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

Shiffrar, M., Lichtey, L., & Heptulla Chatterjee, S. (1997). The perception of biological motion across apertures. *Perception & Psychophysics*, *59*(1), 51–59.

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

Swets, J. A. (1964). *Signal detection and recognition in human observers: Contemporary readings*. New York: John Wiley and Sons.

Tanner, W. P., Jr., & Birdsall, T. G. (1958). Definitions of d′ and η as psychological measures. *The Journal of the Acoustical Society of America*, *30*(10), 9227–928.

Thompson, B., Hansen, B. C., Hess, R. F., & Troje, N. F. (2007). Peripheral vision: Good for biological motion, bad for signal noise segregation? *Journal of Vision*, *7*(10):12, 1–17, doi:10.1167/7.10.12. [PubMed] [Article]

Thurman, S. M., & Lu, H. (2013a). Complex interactions between spatial, orientation, and motion cues for biological motion perception across visual space. *Journal of Vision*, *13*(2):8, 1–18, doi:10.1167/13.2.8. [PubMed] [Article]

Thurman, S. M., & Lu, H. (2013b). Physical and biological constraints govern perceived animacy of scrambled human forms. *Psychological Science*, *24*(7), 1133–1141, doi:10.1177/0956797612467212.

Tjan, B. S., Braje, W. L., Legge, G. E., & Kersten, D. (1995). Human efficiency for recognizing 3-D objects in luminance noise. *Vision Research*, *35*(21), 3053–3069.

Watson, A. B., & Pelli, D. G. (1983). QUEST: A Bayesian adaptive psychometric method. *Perception & Psychophysics*, *33*(2), 113–120.

Vangeneugden, J., Peelen, M. V., Tadin, D., & Battelli, L. (2014). Distinct neural mechanisms for body form and body motion discriminations. *Journal of Neuroscience*, *34*(2), 574–585.

van Boxtel, J. J., & Lu, H. (2011). Visual search by action category. *Journal of Vision*, *11*(7):19, 1–14, doi:10.1167/11.7.19. [PubMed] [Article]

Appendix: IO analysis

All targets were assumed to be known to the ideal observer. The frame rate of the movies were also assumed to be known. A target was represented by dynamic templates $\{T_i\}$, corresponding to the noise-free image frames of the target, in which $i = 1, \ldots, n$. Each frame had the same image size, consisting of M pixels. Given a motion sequence of n image frames$\{I_i\}$, the probability of a target $\{T_i\}$ being present can be expressed using Bayes rule:

$$P(\{T_i\}|\{I_i\}) = \frac{P(\{I_i\}|\{T_i\})P(\{T_i\})}{P(\{I_i\})}. \quad (1)$$

Detection

Given an observed sequence$\{I_i\}$, detection amounted to deciding between the following two hypotheses: Hypothesis 1 (T): The target was present in $\{I_i\}$, and Hypothesis 2 (N): The target was absent in $\{I_i\}$.

If $P(T|\{I_i\}) > P(N|\{I_i\})$, the target (a walker) was judged to be present. If $P(T|\{I_i\}) < P(N|\{I_i\})$, the target was judged to be absent. The ratio of two posterior probabilities is

$$\frac{P(T|\{I_i\})}{P(N|\{I_i\})} = \frac{P(\{I_i\}|T)}{P(\{I_i\}|N)} \cdot \frac{P(T)}{P(N)}. \quad (2)$$

Because the rightward-walking target was present in half the trials and absent in the other half, the prior probabilities were equal, $\frac{P(T)}{P(N)} = 1$. As a result, the optimal decision was determined by the likelihood ratio, $\frac{P(\{I_i\}|T)}{P(\{I_i\}|N)}$. The likelihood term $P(\{I_i\}|T)$ quantified how likely a given image sequence $\{I_i\}$ was generated by adding Gaussian noise on a target (walker) template T. $P(\{I_i\}|N)$ quantified the probability that $\{I_i\}$ was generated by adding Gaussian noise on a uniform background image, B.

Due to the spatial uncertainty, the IO simulation summed over all possible spatial locations, with τ denoting that the center position of the walker was located within a certain window w. The display location τ was randomly sampled according to a uniform distribution, yielding $P(\tau) = \frac{1}{w_x w_y}$, w_x and w_y denote the size of the window w.

$$\begin{aligned} P(\{I_i\}|T) &= \sum_{\tau \in w} P(\tau) P(\{I_i\}|\{T_i\}, \tau) \\ &= \sum_{\tau \in w} P(\tau) \prod_{i=1}^{n} \frac{1}{(\sqrt{2\pi}\sigma)^M} \exp\left(-\frac{\|I_i - T_{i,\tau}\|^2}{2\sigma^2}\right), \end{aligned} \quad (3)$$

where $T_{i,\tau}$ denotes that the template target is located at the location τ.

The likelihood term $P(\{I_i\}|N)$ quantified how likely a given image sequence $\{I_i\}$ was observed from a noise background, N.

$$\begin{aligned} P(\{I_i\}|N) &= P(\{I_i\}|\{N_i\}) \\ &= \prod_{i=1}^{n} \frac{1}{(\sqrt{2\pi}\sigma)^M} \exp\left(-\frac{\|I_i - N_i\|^2}{2\sigma^2}\right). \end{aligned} \quad (4)$$

Discrimination

Given a walking sequence $\{I_i\}$, discrimination amounted to deciding between the two alternative hypotheses: Hypothesis 1 (R): rightward walking, and Hypothesis 2 (L): leftward walking. If $P(R|\{I_i\}) > P(L|\{I_i\})$, then the walking direction was identified as toward the right and otherwise toward the left. The ratio of the two posterior probabilities was equal to the likelihood ratio, $\frac{P(\{I_i\}|R)}{P(\{I_i\}|L)}$ because the walking direction was toward the left in half of the trials and the right in the other half. The calculation of the likelihood term $P(\{I_i\}|R)$ and $P(\{I_i\}|L)$ amounted to comparing the image sequence $\{I_i\}$ with stored template sequences $\{T_i{}^R\}$ and $\{T_i{}^L\}$, respectively representing rightward and leftward walking.

The stimulus contained a target walker, perturbed by the addition of spatially and temporally independent Gaussian luminance noise with 0 mean contrast and standard deviation σ. The ideal observer simulation summed over all possible spatial locations in which the target could appear,

$$\begin{aligned} P(\{I_i\}|R) &= \sum_{\tau \in w} P(\tau) \prod_{i=1}^{N} P(I_i|T_i{}^R, \tau) \\ &= \sum_{\tau \in w} P(\tau) \prod_{i=1}^{N} \frac{1}{(\sqrt{2\pi}\sigma)^M} \exp\left(-\frac{\|I_i - T_{i,\tau}{}^R\|^2}{2\sigma^2}\right), \end{aligned} \quad (5)$$

$$\begin{aligned} P(\{I_i\}|L) &= \sum_{\tau \in w} P(\tau) \prod_{i=1}^{N} P(I_i|T_i{}^L, \tau) \\ &= \sum_{\tau \in w} P(\tau) \prod_{i=1}^{N} \frac{1}{(\sqrt{2\pi}\sigma)^M} \exp\left(-\frac{\|I_i - T_{i,\tau}{}^L\|^2}{2\sigma^2}\right). \end{aligned} \quad (6)$$

The IO used the following decision rule to make a response:

$$\begin{cases} \text{if } \frac{P(\{I_i\}|R)}{P(\{I_i\}|L)} > 1, & \text{decide rightward walking} \\ \text{if } \frac{P(\{I_i\}|R)}{P(\{I_i\}|L)} < 1, & \text{decide leftward walking} \end{cases}. \quad (7)$$

Definition of efficiency

Statistical efficiency (Fisher, 1925; Swets, 1964) is typically defined as the squared ratio of d's for the IO and human observers:

$$\eta = \left(\frac{d'_{Human}}{d'_{Ideal}}\right)^2, \quad (8)$$

where human and ideal sensitivity d' are measured on stimuli with the same signal contrast at a given noise level. Because sensitivity d' is proportional to the square root of the signal contrast energy, efficiency is also defined as the ratio of signal energy threshold E_t for IOs and human observers to perform a task at a given level of performance (Tanner & Birdsall, 1958). In the present study, we used Equation 9 to compute human efficiency in different tasks:

$$\eta = \frac{E_{t,Ideal}}{E_{t,Human}}. \quad (9)$$