Human Action Classification from Naturalistic Videos

Uriel González-Bravo (uriel.gonzalezbr@rutgers.edu)
Dept. of Psychology, 152 Frelinghuysen Road Piscataway, NJ 08854, USA

Jacob Feldman (jacob@ruccs.rutgers.edu)
Dept. of Psychology, Center for Cognitive Science, 152 Frelinghuysen Road Piscataway, NJ 08854, USA

Abstract

It has long been known that human observers can identify actions based on how people move, even from very impoverished motion depictions such as Point Light Displays (PLDs). This study investigates how humans classify actions, and what types of motion information they use to do so. Using a newly available technique (OpenPose) for extracting human joint locations from natural video, we created three types of reduced displays: PLDs, stick figures, and motion flow videos. Participants identified actions in these videos through verbal responses, and these responses were analyzed for semantic similarity using a Natural Language Processing model. A Hierarchical Bayesian Model further compared semantic similarities across video conditions. Results showed the highest intersubjective agreement (a proxy for proportion correct) for stick figures, followed by PLDs, and the lowest for motion flow videos. These results suggest that dynamic pose representations are crucial for accurate action classification, with motion flow supporting only coarse classification. The same pattern held across different action categories, such as instrumental versus locomotion and upper versus lower limb actions.

Keywords: Biological motion; action classification; visual perception; bayesian modeling

Introduction

Classifying human action based on movement is an important everyday task. We can easily tell if someone is running or walking, brushing their teeth or combing their hair, or simply sitting and reading. Indeed human observers can infer intentions, emotions, or motivations from human movement (see Cutting, 2013). This ability extends beyond human figures to inanimate objects such as geometric shapes (see Heider & Simmel, 1944; Epley, Waytz, & Cacioppo, 2007). But the underlying mechanisms of action classification are not well understood.

Biological motion

Johansson (1973) introduced a class of stimuli, now usually called "Point Light Displays" (PLDs), which opened a window onto how the brain interprets human movement. Originally, PLDs were created by attaching lights to the joints of an actor wearing a black suit, and filming them in the dark so only the movement of joints (ankles, wrists, knees, elbows, and shoulders) could be seen. Even with such impoverished stimuli, human observers can still recognize simple actions such as walking, running, and dancing. On the basis of PLDs humans can also perceived intentions (Okruszek, 2018), identify gender (Mather & Murdoch, 1994), discern emotions (Dittrich, Troscianko, Lea, & Morgan, 1996; Parkinson, Walker, Memmi, & Wheatley, 2017; Ikeda, Destler, & Feldman, 2025), and even recognize individuals (Troje, Westhoff, & Lavrov, 2005). The ability to interpret PLDs is substantially cross-cultural; for example Pica, Jackson, Blake, and Troje (2011) showed that the Mundurucu, a culturally isolated group in Brazil, can identify point-light walkers in coherent, inverted, and spatially and temporally scrambled conditions. These days the term "biological motion" is usually associated with the study of PLDs, but note that the mechanisms by which people interpret the movements of living things can be studied using a variety of stimulus types.

Motion flow and skeletal models of shape

What visual features does the brain use to interpret biological motion? Since Johannson's original studies, it is often been assumed that action interpretation rests on the perceptual organization of the stimulus into a human form. But conversely, some authors have proposed multiple streams of processing, often one based on form information and another based on motion flow. These proposals are related to the well-known distinction between dorsal and ventral streams of processing (see Lichtensteiger, Loenneker, Bucher, Martin, & Klaver, 2008). The dorsal stream locates and tracks objects, supporting spatial orientation and movement coordination, while the ventral stream identifies and recognizes objects. In the realm of biological motion perception, some authors (e.g., Mather, Pavan, Bellacosa Marotti, Campana, & Casco, 2013, and Misaghian, Lugo, & Faubert, 2022) have suggested that the dorsal system employs motion flow to distinguish biological movements from generic motion flow patterns, while the ventral system facilitates the estimation of poses, and thus enables more detailed interpretations of specific actions.

Along these lines, Giese and Poggio (2003) proposed a hierarchical neural model that integrates form and motion pathways to interpret biological motion. This model utilizes two pathways: one starting from local orientation detectors and supporting form representation, and the other starting from motion detectors and tracking movement. Extending this logic, Casile and Giese (2005) demonstrated that motion flow, corresponding roughly to the dorsal stream, suffices to

In D. Barner, N.R. Bramley, A. Ruggeri and C.M. Walker (Eds.), *Proceedings of the 47th Annual Conference of the Cognitive Science Society*

allow some recognition of some actions. They used a novel point-light stimulus, the "critical features stimulus," which contained dominant motion features combined with coarse positional information. These stimuli were crafted to isolate specific motion features that might be essential for biological motion perception. A more recent model that follows a similar path was proposed by Misaghian et al. (2022), who introduced a risk-averse Bayesian approach placing special emphasis on the dorsal pathway for detecting biological motion.

In the ventral component of these models, as in the original Johansson framework, action classification is supported by integrated representations of the human form. The nature of those representations is however not completely clear. One important proposal for how biological forms can be represented is Blum's (1973) medial axis transform and its descendants (Burbeck & Pizer, 1995; Kimia, 2003; Feldman & Singh, 2006; Ayzenberg & Lourenco, 2019). These skeletal representations center on the "bones" of perceptual shapes, in Blum's original proposal local symmetry axes, which tend to correspond roughly to limbs and other components of biological forms (see Kovács, Fehér, & Julesz, 1998). More recent skeletal frameworks have conceptualized the axes somewhat differently from Blum, but all share the goal of decomposing shapes into individual parts whose spatial relations define the object's "pose." Burbeck and Pizer (1995) suggested that that neurons in the visual system identify shapes using axis-like structures, or "cores," described as being equidistant from the nearest contour boundaries. Feldman and Singh (2006) proposed a Bayesian framework for estimating axial structure (see also Wilder, Feldman, & Singh, 2011; Destler, Singh, & Feldman, 2023), emphasizing the goal of a one-to-one mapping between estimated axes and coherently articulating parts. Most of these proposals are based on static stimuli, and the connection between skeletal shape representations and biological motion has scarcely been explored. The representation of nonrigidly deforming shape is itself very poorly understood, as the vast majority of the structure-from-motion literature assumes rigidly moving objects (see Choi, Feldman, & Singh, 2024). Hence there is a substantial gap in the literature concerning the role of skeletal representations in the interpretation of biological motion.

Limitations in biological motion research

Notwithstanding the huge literature on biological motion, most studies have focused on a very limited number of actions, most often walking and running. Johansson's own studies (1973, 1976) only mention a few actions. This limitation poses significant challenges when attempting to draw broader generalizitions. Many studies of the interpretation of PLDs use only walking (e.g. Rutherford & Kuhlmeier, 2013). Even more critically, the claim that motion flow alone might be sufficient to recognize actions is usually based on very small number of actions, sometimes just walking, with performance gauged by the subject's ability to distinguish walking from random motion patterns.

Some studies have used a broader range of actions. Dittrich (1993) used 12 different actions divided into three categories: locomotive (walking, going upstairs, jumping, leaping), instrumental (hammering, ball bouncing, stirring, and box lifting), and social (dancing, boxing, greeting, and threatening). He found that locomotive actions were easier to recognize than instrumental and social ones. Similarly, van Boxtel and Lu (2011), using 38 animations spread across four categories (boxing, dancing, walking, and running). A hierarchical clustering analysis of search slopes drawn from a visual search task yielded a fairly intuitive clustering of action types. Walking and running, for example, share more common features than walking and boxing. Moreover, actions that are easier to group and generalize are typically locomotive, such as running and walking. In contrast, this generalization is less likely with non-locomotive actions, like boxing or knee bends (cf. Giese & Lappe, 2002).

The primary obstacle to using a large range of actions in biological motion studies is the difficulty of producing suitable videos, which often requires special equipment or techniques (see Ghorbani et al., 2021). To avoid this limitation, in the current study, we employ an alternative methodology: OpenPose (Cao, Simon, Wei, & Sheikh, 2017). This framework uses advanced deep learning techniques, specifically Convolutional Neural Networks (CNNs), to detect and track human joint points in real time directly from videos. OpenPose enables highly accurate real-time pose estimation, even in situations where people overlap or interact with each other (Washabaugh, Shanmugam, Ranganathan, & Krishnan, 2022).

In what follows, we use OpenPose to create displays from a much wider range of naturalistic actions than would be possible with more conventional methods. Our action classes cover a huge span of action types, from simple locomotion (walking, running) to dancing, sports, and a range of more precise actions such as brushing teeth or combing hair (see Fig. 5 for the complete list). In the traditional biological motion literature, one often encounters broad summaries such as "people can recognize actions accurately." But given the very limited range of actions usually tested, this summary is almost certainly oversimplified. As Dittrich (1993) found, some actions are recognized much more easily than others. By using a much broader range of natural actions, our hope is to be able to quantify the human ability to classify actions more comprehensively.

In addition, we aim to study which specific motion features support action classification by presenting human forms in several different ways. The results of OpenPose are often depicted as a "stick figure" for each human form detected. However the procedure actually outputs joint positions, which for the purposes of parametric experiments can be

depicted in stimuli in several different ways. For example, if each joint position is rendered as a dot but *not* connected to other joints, the result is an automatically generated PLD. One of the hypotheses underlying the following study is that if human action classification rests specifically on skeletal representations, then stick figures (which make skeletal information overt in the stimulus) should support better action classification than PLDs (which do not).

Moreover, in the studies below we include stimuli depicting only the motion flow from the same videos, rendered via random dots with motion flow matching that in the raw video. This condition allows us to measure action classification from motion flow alone, as speculated to be accomplished by the dorsal stream without benefit of human form integration. By applying this presentation method to a broad range of naturalistic actions, we hope to give a more comprehensive answer to the question of whether, or to what degree, action can be classified on the basis of motion alone.

With that in mind, our study has three aims. First, we aim to utilize a broader variety of actions drawn from naturalistic video, in order to more comprehensively assess human observers' ability to classify human action. Second, by comparing different methods of depiction (PLDs vs. stick figures) we aim to better understand the importance of skeletal representations in action classification. Third, by comparing these two methods (both of which allow the human form to be integrated, albeit more directly from stick figures than PLDs) to motion flow videos, we hope to assess the degree to which action can be classified from motion flow alone.

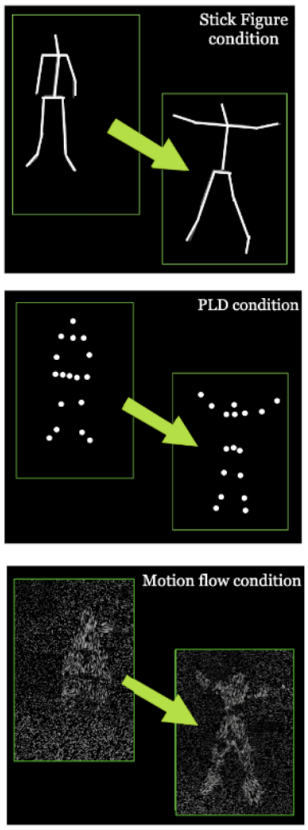

Figure 1: Stick figure (top), PLD (middle), and motion flow display (bottom).

Methods

Experimental framework

Forty-six students from Rutgers University were recruited for this study and were given credits for their Psychology class for their participation. We used 78 actions, including everyday activities like brushing teeth, drinking water, walking, etc. (see Fig. 5). For these 78 actions, three distinct types of videos were created: PLDs (consisting of dots at joint locations), stick figures (joint positions joined by line segments in an anatomically correct body plan), and motion flow videos (flow fields estimated using the Lucas-Kanade algorithm, Patel & Upadhyay, 2013). Fig. 1 illustrates the three conditions. The 78 videos were divided into three groups of 26 each, with each group presented in one condition to each subject, ensuring that no subject ever saw the same action twice. The assignment of groups to conditions was then permuted across three groups of participants so that across subjects all actions were presented an equal number of times in each condition. The subjects' task was to freely describe in a text box the action depicted in each video.

As a dependent variable, we focus on the *intersubjective agreement rate* θ among participants' responses, calculated using Natural Language Processing (NLP) as explained below. The goal of using intersubjective agreement rather than proportion correct, as is more conventional, is to avoid stipulating correct classifications for each video (in most studies, provided intuitively by the experimenters themselves). Previous studies of action interpretation (e.g. Ikeda et al., 2025) have found that intersubjective agreement is highly correlated with proportion correct (i.e., proportion agreement with human ratings of the raw videos by a separate subject group). Intersubjective agreement has the advantage of not requiring that ground truth ratings be either stipulated or collected from other subjects, but can be calculated using only the subjects' responses to the stimuli themselves.

To automatically compute intersubjective agreement, we used a pre-trained English model *en_core_web_md* from SpaCy in Python to transform the participants' text responses into semantic vectors (see Honnibal, Montani, Landeghem, & Boyd, 2020). Then the cosine similarity between subjects' response vectors was calculated for each video, providing a quantitative measure of how similar the responses were to each other in terms of semantic content. The cosine similarity ranges from 0 to 1, with higher values indicate

greater similarity. In the analyses below, θ denotes the average value of these similarities.

Bayesian model

To analyzes these data, we constructed a hierarchical Bayesian model of agreement rate among participants. The purpose of this model is to quantify the influence of the conditions, in particular display condition, on the intersubjective similarity measure θ. By using this approach, we can account for the hierarchical structure of our data, where responses are nested within participants, and participants are nested within video types. This model allows us to estimate the influence of each display type on the semantic similarity of action descriptions while considering individual differences among participants.

The model estimates posterior distributions for the effects of display types on θ, providing a comprehensive view of the probable values of these effects in light of the data. Figure 2 give a graphical representation of the model. In the figure, as is standard, circles represent continuous variables, squares represent discrete variables, filled shapes denote known variables, and empty shapes are parameters to be estimated. For instance, θ_{ijk} is the agreement rate calculated for participant i for video j under condition k, computed using the cosine similarity method. It is assumed that θ_{ijk} follows a Beta distribution with parameters α_{jk} and β_{jk}, indicating that each video j in condition k has its unique distribution. The choice of a Beta distribution for the agreement rate is due to its values being constrained between 0 and 1, making it a natural choice for modeling probabilities. Additionally, the Beta distribution is a standard choice in Bayesian modeling for probability parameters, in part because it is flexible and can accommodate a wide range of distributional forms, and also because it is conjugate to the binomial distribution (see Lambert, 2018). Finally, the α_{jk} and β_{jk} parameters of the Beta distribution each follow Gamma distributions parameterized by (τ_k, ν_k) and (γ_k, κ_k), respectively, which means that each condition has its own parameter structure. These hyperparameters $\tau_k, \nu_k, \gamma_k, \kappa_k$ themselves follow Gamma distributions with parameters (0.001, 0.001), ensuring a weakly informative prior that encourages learning from the data while maintaining flexibility in the model (Lambert, 2018).

Results

Figure 3 shows the posterior distribution of the agreement rate θ, broken down by display type. Stick figures show the highest agreement rate, followed PLDs condition, and motion flow the lowest. However agreement rates for motion flow were still far above zero, with over 99% of the area under the curve lying above zero. That is, though performance in the motion flow condition is generally the worst, many actions

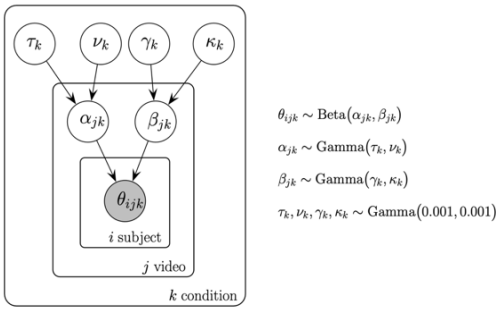

Figure 2: Hierarchical Bayesian Model

are indeed recognizable from motion alone.

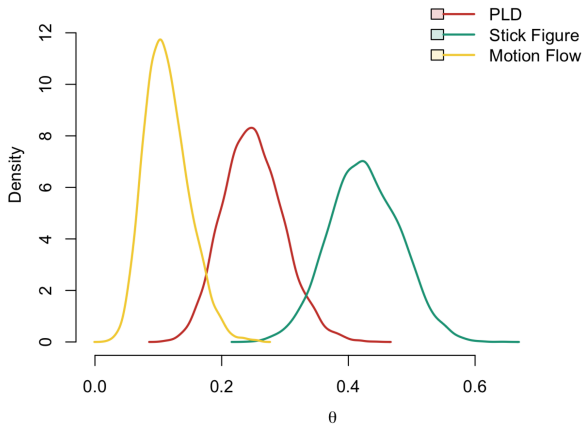

Figure 3: Posterior distribution over θ, the intersubjective agreement rate, broken down by display condition

To compare posterior distributions, we used Kolmogorov-Smirnov (KS) distances. The KS distance can range from 0 to 1, where zero indicates that the distributions are identical, and one indicates the maximum possible difference between them. The KS distances results showed significant differences for all comparisons ($p < 0.01$): stick figure vs. motion flow at 0.9978, PLDs vs. Flow at 0.8920, and stick figure vs. PLDs at 0.8856. In addition, we performed a Bayes Factor (BF) analysis to compare hypotheses H_1, the distributions are different, and H_0, the distributions are the same. The BF_{10} values were: stick figures vs. PLDs, 2.26×10^{3262}; PLDs vs. motion Flow, 1.09×10^{3662}; and stick figures vs. motion flow: 6.37×10^{6616}. These results strongly support the conclusion that these distributions are indeed all different from each other.

Next we conducted two analyses to understand which types of actions were more and less interpretable. First, following Dittrich (1993), we divide the videos into *locomotive* vs. *instrumental* actions. Instrumental actions involve movements aimed at manipulating objects or performing

tasks that require interaction with tools or equipment, such as lifting an object, dribbling a basketball, or flipping a coin. These actions are generally goal-directed and involve the use of hands or other body parts to accomplish a specific task. Locomotive action are those aimed at changing location or position, such as walking, jumping, or dancing. These actions primarily use the legs and focus on transportation or general movement.

Figure 4a shows posterior distributions over intersubjective agreement for locomotive vs. instrumental actions, again broken down by display type. First, we found the same performance ordering (stick figures > PLDs > motion flow) within both locomotive and instrumental actions. Corresponding values for KS distances and BFs can be found in Table 1. As observed in Fig. 4b, generally locomotive actions are more recognizable than instrumental actions, and the KS differences among display types are all consistent and statistically substantial ($p < 0.01$): 0.244 for stick figure differences, 0.212 for PLDs, and 0.10614 for flow videos. For the Bayes Factor analysis, our H_1 hypothesis was that the distributions are different, while our H_0 hypothesis was that they are the same. The BF_{10} values obtained were 4.3×10^{3149} for the Stick Figure videos, 2.3×10^{3104} for the PLDs, and 1.07×10^{777} for the motion flow videos. Fig. 5 shows mean intersubjective agreeement for all actions, averaging over display type.

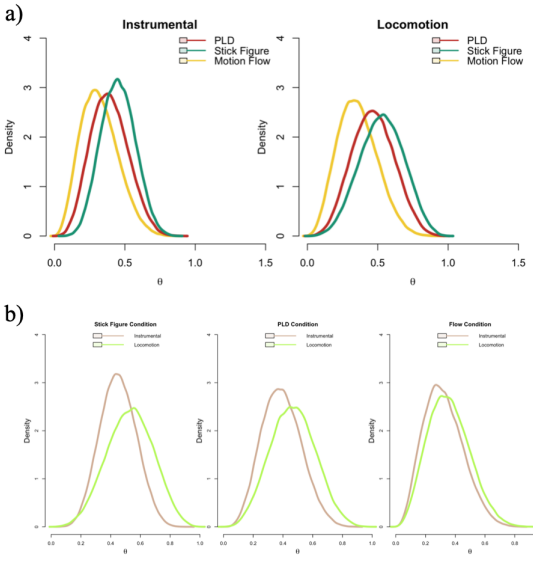

Figure 4: Instrumental and Locomotion Posterior Distributions

		KS Value	BF_{10}
Instrumental	stick F.-PLDs	0.1967^{***}	1.75×10^{2382}
	PLDs-Flow	0.2111^{***}	1.30×10^{2927}
	stick F.-Flow	0.4021^{***}	3.04×10^{10171}
Locomotion	stick F.-PLDs	0.16759^{***}	1.79×10^{1714}
	PLDs-Flow	0.29707^{***}	2.37×10^{5846}
	stick F.-Flow	0.44065^{***}	3.30×10^{12589}

Table 1: Comparison of KS values and BF_{10} for Instrumental and Locomotion actions. *** indicates $p < 0.01$.

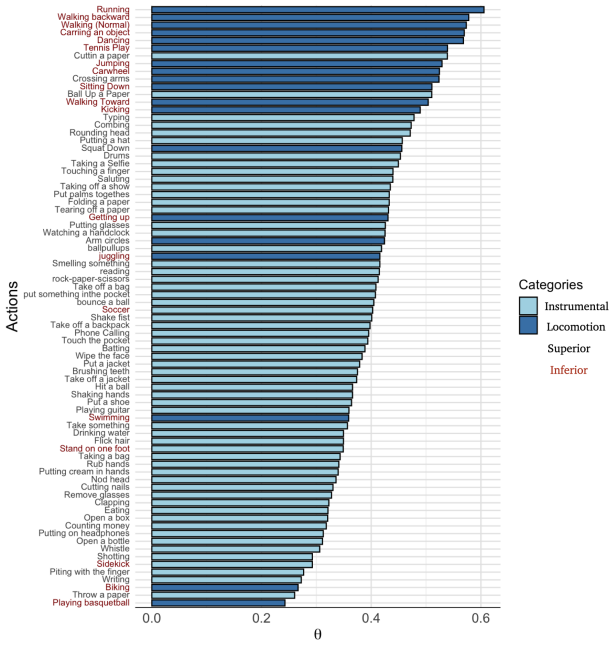

Figure 5: Intersubjective agreement for all 78 actions. Bar color indicates locomotive vs. instrumental action; text color indicates superior vs. inferior motion.

As noted by Dittrich (1993), locomotive actions generally involve lower limbs and instrumental actions involve upper limbs. Our second analysis focuses more specifically on this distinction. We divided our videos into two types: actions primarily involving the upper body, such as the head or arms, which we refer to as *superior*, and actions primarily involving the lower body, such as the feet or legs, which we refer to as *inferior*. This classification was determined by calculating the average displacement of joint points detected by OpenPose. We then grouped the videos based on the frequency of joint position changes from frame to frame. As expected, inferior actions show greater intersubjective agreement than superior actions (Figure 6a), with the same performance ordering among display types (stick figures > PLDs > motion flow videos). As in our previous analyses, we calculated the KS distances from the posterior distributions to quantify the difference between the videos based on upper and lower limbs and the Bayes Factor to confirm if the distributions differ (Table 2).

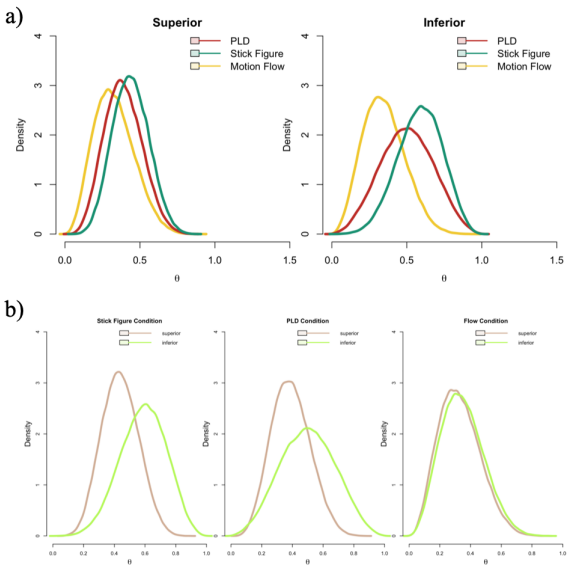

Figure 6: Posterior Distributions of Superior and Inferior Limb Actions

		KS Value	BF_{10}
Instrumental	stick F.-PLDs	0.1967^{***}	1.75×10^{2382}
	PLDs-Flow	0.2111^{***}	1.30×10^{2927}
	Stick F.-Flow	0.4021^{***}	3.04×10^{10171}
Locomotion	stick F.-PLDs	0.16759^{***}	1.79×10^{1714}
	PLDs-Flow	0.29707^{***}	2.37×10^{5846}
	stick F.-Flow	0.44065^{***}	3.30×10^{12589}

Table 2: Comparison of KS values and BF_{10} for Instrumental and Locomotion actions. The three asterisks (***) next to each KS value indicate $p < 0.01$.

As shown in Figure 6b, the agreement rate is generally higher for inferior (lower-limb) actions. This difference is particularly notable in the stick figure condition, and is somewhat reduced in motion flow condition. To quantify the differences between the posterior distributions of each video type for upper and lower limbs, we calculated the KS distances, obtaining values of 0.42243 ($p < 0.01$) for stick figures, 0.30907 ($p < 0.01$) for PLDs, and 0.05463 ($p < 0.01$) for motion flow videos. To confirm these results, we performed Bayes Factor analysis where H_1 posits that the distributions are different and H_0 posits that they are the same. Our results parallel the previous analyses: we obtained a BF_{10} of 6.1×10^{11379} for stick figures, 7.1×10^{5593} for PLDs, and 6.85×10^{193} for motion flow. In summary, actions involving the lower body are easier to recognize however than those involving the upper body, though in the motion flow condition this advantage is greatly reduced.

Summary and conclusions

This study examined how people classify human action from motion, using an unprecedentedly wide range of action categories, depicted in several different ways. Using a Hierarchical Bayesian Model, the results consistently showed that stick figure videos elicited the best performance (highest intersubjective agreement), followed by Point-Light Displays (PLDs), and then motion flow videos. This pattern was observed across different action categories, such as instrumental versus locomotion actions and upper versus lower limb movements.

These findings support several long-held assumptions about the mechanisms of human action classification, and also entail several novel conclusions.

First, as Johannson had originally speculated, action classification rests heavily on the integration of the human form, which enables a precise representation of the dynamically changing human pose. This is why stick figures and traditional PLDs (both of which allow body pose to be accurately estimated) support better performance than the motion flow condition (from which body pose is difficult to estimate). We observed this difference over a large range of different action types, and it recurred in several different subclassifications of motion such as locomotion and instrumental action.

In addition, as originally suggested by Giese and Poggio (2003) and Casile and Giese (2005), action classification seems to proceed somewhat differently along dorsal and ventral streams, and as a result subjects were able to perform well above chance even in the motion flow condition, presumably primarily on the basis of the dorsal stream. From our data, action classification from motion flow alone is comparatively coarse, and probably would not support the sort of precise comprehension of others' actions that humans exhibit in everyday life.

But several novel conclusions emerge as well. First, the consistent superiority in performance in the Stick Figure condition relative to PLDs suggests that the form integration that supports action classification depends specifically on *skeletal* representations of human form. This hypothesis emerges from the literature on static shape representations, but has not to our knowledge previously been tested on dynamic stimuli. As Kovács et al. (1998) had speculated, one of the main benefits of skeletal representations is their suitability for representing dynamic articulation in biological forms. Our data suggests that this advantage translates into a measurable advantage in action classification when the underlying skeletal structure is made overt in the stimulus.

Finally, by studying a much larger range of natural action types than previously undertaken, our study demonstrates that the human ability to classify action from dynamic shape alone is quite variable. We can do this very well for some actions (e.g. walking and running, the sort of locomotive action that Johnannson originally emphasized) but much more poorly for more upper-body or instrumental actions (see Fig. 5). This finding opens the door to more specific future studies of human action classification.

References

Ayzenberg, V., & Lourenco, S. F. (2019). Skeletal descriptions of shape provide unique perceptual information for object recognition. *Scientific Reports*, *9*(1), 9359.

Blum, H. (1973). Biological shape and visual science (part I). *Journal of Theoretical Biology*, *38*, 205–287.

Burbeck, C. A., & Pizer, S. M. (1995). Object representation by cores: Identifying and representing primitive spatial regions. *Vision Research*, *35*(13), 1917–1930.

Cao, Z., Simon, T., Wei, S. E., & Sheikh, Y. (2017). Realtime multi-person 2D pose estimation using part affinity fields. In *2017 IEEE conference on computer vision and pattern recognition (CVPR)* (p. 1302-1310). Honolulu, HI, USA. doi: 10.1109/CVPR.2017.143

Casile, A., & Giese, M. A. (2005). Critical features for the recognition of biological motion. *Journal of Vision*, *5*(4), 348-360. doi: 10.1167/5.4.6

Choi, R., Feldman, J., & Singh, M. (2024). Perceptual biases in the interpretation of non-rigid shape transformations from motion. *Vision*, *8*(3).

Cutting, J. E. (2013). Gunnar johansson, events, and biological motion. In K. L. Johnson & M. Shiffrar (Eds.), *People watching: Social, perceptual, and neurophysiological studies of body perception* (p. 11-24). New York: Oxford University Press.

Destler, N., Singh, M., & Feldman, J. (2023). Skeleton-based shape similarity. *Psychological Review*, *130*(6), 1653–1671. Retrieved from https://doi.org/10.1037/rev0000412 doi: 10.1037/rev0000412

Dittrich, W. H. (1993). Action categories and the perception of biological motion. *Perception*, *22*(1), 15–22. Retrieved from https://doi.org/10.1068/p220015 doi: 10.1068/p220015

Dittrich, W. H., Troscianko, T., Lea, S. E. G., & Morgan, D. (1996). Perception of emotion from dynamic point-light displays represented in dance. *Perception*, *25*, 727–738.

Epley, N., Waytz, A., & Cacioppo, J. (2007). On seeing human: A three-factor theory of anthropomorphism. *Psychological Review*, *114*, 864-886.

Feldman, J., & Singh, M. (2006). Bayesian estimation of the shape skeleton. *Proceedings of the National Academy of Sciences*, *103*(47), 18014–18019.

Ghorbani, S., Mahdaviani, K., Thaler, A., Kording, K., Cook, D., Blohm, G., et al. (2021). MoVi: A large multi-purpose human motion and video dataset. *PLoS ONE*, *16*(6), e0253157. Retrieved from https://doi.org/10.1371/journal.pone.0253157 doi: 10.1371/journal.pone.0253157

Giese, M. A., & Lappe, M. (2002). Measurement of generalization fields for the recognition of biological motion. *Vision Research*, *42*(15), 1847–1858. doi: 10.1016/S0042-6989(02)00093-7

Giese, M. A., & Poggio, T. (2003). Neural mechanisms for the recognition of biological movements. *Nature Reviews Neuroscience*, *4*, 179-192. doi: 10.1038/nrn1057

Heider, F., & Simmel, M. (1944). An experimental study of apparent behavior. *The American Journal of Psychology*, *57*, 243-259.

Honnibal, M., Montani, I., Landeghem, S. V., & Boyd, A. (2020). *spacy: Industrial-strength natural language processing in python.* https://doi.org/10.5281/zenodo.1212303. (Accessed: 2025-05-10)

Ikeda, E., Destler, N., & Feldman, J. (2025). The role of dynamic shape cues in the recognition of emotion from naturalistic body motion. *Attention, Perception, & Psychophysics*, 1–15.

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

Johansson, G. (1976). Spatio-temporal differentiation and integration in visual motion perception. *Psychological Research*, *38*, 379–393. Retrieved from https://doi.org/10.1007/BF00309043 doi: 10.1007/BF00309043

Kimia, B. B. (2003). One the role of medial geometry in human vision. *Journal of Physiology Paris*, *97*, 155–190.

Kovács, I., Fehér, Á., & Julesz, B. (1998). Medial-point description of shape: A representation for action coding and its psychophysical correlates. *Vision Research*, *38*(15-16), 2323–2333. Retrieved from https://doi.org/10.1016/S0042-6989(97)00321-0 doi: 10.1016/S0042-6989(97)00321-0

Lambert, B. (2018). *A student's guide to bayesian statistics* (illustrated ed.). SAGE Publications.

Lichtensteiger, J., Loenneker, T., Bucher, K., Martin, E., & Klaver, P. (2008). Role of dorsal and ventral stream development in biological motion perception. *NeuroReport*, *19*(18), 1763–1767. doi: 10.1097/WNR.0b013e328318ede3

Mather, G., & Murdoch, L. (1994). Gender discrimination in biological motion displays based on dynamic cues. *Proceedings: Biological Sciences*, *258*(1353), 273–279.

Mather, G., Pavan, A., Bellacosa Marotti, R., Campana, G., & Casco, C. (2013). Interactions between motion and form processing in the human visual system. *Frontiers in Computational Neuroscience*, *7*, 65.

Misaghian, K., Lugo, J. E., & Faubert, J. (2022). Descriptive risk-averse bayesian decision-making, a model for complex biological motion perception in the human dorsal pathway. *Biomimetics*, *7*, 193. doi: 10.3390/biomimetics7040193

Okruszek, L. (2018). It is not just in faces! processing of emotion and intention from biological motion in psychiatric disorders. *Frontiers in Human Neuroscience*, *12*, 329867. Retrieved from https://doi.org/10.3389/fnhum.2018.00048

Parkinson, C., Walker, T., Memmi, S., & Wheatley, T. (2017).

Emotions are understood from biological motion across remote cultures. *Emotion*, *17*(3), 459–477. Retrieved from https://doi.org/10.1037/emo0000194 doi: 10.1037/emo0000194

Patel, D., & Upadhyay, S. (2013). Optical flow measurement using Lucas Kanade method. *International Journal of Computer Applications*, *61*(10), 6–9.

Pica, P., Jackson, S., Blake, R., & Troje, N. (2011). Comparing biological motion in two distinct human societies. *PLoS ONE*, *6*(12), e28391.

Rutherford, M. D., & Kuhlmeier, V. A. (2013). *Social perception: Detection and interpretation of animacy, agency, and intention* (M. D. Rutherford & V. A. Kuhlmeier, Eds.). Boston Review. doi: 10.7551/mitpress/9780262019279.001.0001

Troje, N., Westhoff, C., & Lavrov, M. (2005). Person identification from biological motion: effects of structural and kinematic cues. *Percept Psychophys*, *67*(4), 667–675. doi: 10.3758/bf03193523

van Boxtel, J. J. A., & Lu, H. (2011). Visual search by action category. *Journal of Vision*, *11*(7), 19. Retrieved from https://doi.org/10.1167/11.7.19 doi: 10.1167/11.7.19

Washabaugh, E. P., Shanmugam, T. A., Ranganathan, R., & Krishnan, C. (2022). Comparing the accuracy of open-source pose estimation methods for measuring gait kinematics. *Gait & Posture*, *97*, 188–195. doi: 10.1016/j.gaitpost.2022.08.008

Wilder, J., Feldman, J., & Singh, M. (2011). Superordinate shape classification using natural shape statistics. *Cognition*, *119*(3), 325–340. Retrieved from https://doi.org/10.1016/j.cognition.2011.01.009 doi: 10.1016/j.cognition.2011.01.009