Kinematic Primitives in Action Similarity Judgments: A Human-Centered Computational Model

Vipul Nair, Paul Hemeren, *Member, IEEE*, Alessia Vignolo, Nicoletta Noceti, Elena Nicora, Alessandra Sciutti, *Member, IEEE*, Francesco Rea, Erik Billing, Mehul Bhatt, Francesca Odone, and Giulio Sandini, *Member, IEEE*

***Abstract*—This article investigates the role that kinematic features play in human action similarity judgments. The results of three experiments with human participants are compared with the computational model that solves the same task. The chosen model has its roots in developmental robotics and performs action classification based on learned kinematic primitives. The comparative experimental results show that both model and human participants can reliably identify whether two actions are the same or not. Specifically, most of the given actions could be similarity judged based on very limited information from a single feature domain (velocity or spatial). Both velocity and spatial features were however necessary to reach a level of human performance on evaluated actions. The experimental results also show that human performance on an action identification task indicated that they clearly relied on kinematic information rather than on action semantics. The results show that both the model and human performance are highly accurate in an action similarity task based on kinematic-level features, which can provide an essential basis for classifying human actions.**

***Index Terms*—Action matching, action similarity, biological motion, comparative study, computational model, kinematic primitives, optical flow (OF), point light display.**

Manuscript received 5 February 2022; revised 10 July 2022 and 12 October 2022; accepted 14 January 2023. Date of publication 30 January 2023; date of current version 11 December 2023. This work was supported in part by the Knowledge Foundation, Stockholm, through SIDUS under Grant 20140220 (AIR, Action and Intention Recognition in Human Interaction With Autonomous Systems), and in part by AFOSR under Grant FA8655-20-1-7035, and research collaboration between the University of Skövde and the Istituto Italiano di Tecnologia, Genoa. The work of Alessandra Sciutti was supported by the Starting Grant from the European Research Council (ERC) through the European Union's Horizon 2020 Research and Innovation Programme under Grant 804388, wHiSPER. *(Corresponding author: Vipul Nair.)*

This work involved human subjects or animals in its research. The authors confirm that all human/animal subject research procedures and protocols are exempt from review board approval.

Vipul Nair, Paul Hemeren, and Erik Billing are with the School of Informatics, University of Skövde, 541 28 Skövde, Sweden (e-mail: vipul.nair@his.se; paul.hemeren@his.se; erik.billing@his.se).

Alessia Vignolo and Alessandra Sciutti are with the CONTACT Unit, Istituto Italiano di Tecnologia, 16152 Genoa, Italy (e-mail: alessia.vignolo@gmail.com; alessandra.sciutti@iit.it).

Nicoletta Noceti, Elena Nicora, and Francesca Odone are with the MaLGa Center—DIBRIS, Università di Genova, 16146 Genoa, Italy (e-mail: nicoletta.noceti@unige.it; elena.nicora@unige.it; francesca.odone@unige.it).

Francesco Rea and Giulio Sandini are with the RBCS Unit, Istituto Italiano di Tecnologia, 16152 Genoa, Italy (e-mail: francesco.rea@iit.it; giulio.sandini@iit.it).

Mehul Bhatt is with the School of Science and Technology, Örebro University, 702 81 Örebro, Sweden (e-mail: mehul.bhatt@oru.se).

Color versions of one or more figures in this article are available at https://doi.org/10.1109/TCDS.2023.3240302.

Digital Object Identifier 10.1109/TCDS.2023.3240302

I. Introduction

HUMAN vision is highly sensitive to the biological motion patterns created by the movement of other individuals (e.g., [1] and [2]). From a developmental point of view, this sensitivity to visual preferences exists in newborns [3] and increases significantly over the course of 3–24 months [4], [5]. Learning to distinguish between different types of action and action exemplars reflects this sensitivity and visual preferences [6].

Judging action similarity is an essential part of learning action categories and a step toward action understanding. In fact, in most behavioral studies, action similarity has been addressed as a form of measure to understand action semantics [7], action prototypes [8], and imitation [9]. Whereas in computational studies, action similarity is addressed in a similar form under the broad umbrella of human action recognition (HAR). HAR is an area where different data modalities provide different computational models to produce reliable recognition [10].

Judging action similarity is a critical factor in the computational domain with very many application areas, such as in sports analysis [11], [12], human–robot interaction [13], [14], autonomous driving [15], education [16], health monitoring [17], and video surveillance [18]. Action similarity can be complicated in a realistic setting, such as the ambiguity of the action class in multiclass action recognition [19]. To address this problem of ambiguity, the labeling of similarity in action (same or different) was first introduced by Kliper-Gross et al. [20] as a critical task in action recognition. According to Kliper-Gross et al. [20], action similarity labeling aims to determine whether the actors in two video sequences are performing the same or different actions. Labeling algorithms depend primarily on creating a suitable metric for the differences between the actions of the extracted kinematic features (see [19] for a detailed review of the approaches). Kliper-Gross et al. showed a considerable gap (around 65%) between state-of-the-art methods and the success rate of humans on action similarity labeling and argued toward a

principled understanding of what makes actions similar or different [20]. The work presented in this article attempts to reduce this gap by investigating the role of kinematic primitives [21], [22] in judging action similarity for humans in relation to a chosen computational model.

In terms of the modality of the action, we specifically pick two types of modalities—optical flow (OF) from the red, green, and blue (RGB) color space and 3-D skeleton data. Both modalities are widely used in research with single-modality models within HAR [10], [23], [24]. Numerous studies directly use or incorporate OF in their computation, for example, to detect moving regions [25], and to track the trajectory of moving objects [26]. Similarly, 3-D skeleton data from either sensors or markers are incorporated in HAR frameworks for several purposes such as calculating joint relations [27] and calculating the temporal variance of joints [28]. Furthermore, 3-D data provides point-light displays (PLDs) which are ideal for testing human action judgment based on kinematics. PLDs remove contextual information and thereby allow a fairer comparison with the model performance. For these reasons, these two modalities are used to investigate action similarity judgments.

This article addresses three specific questions.

1) To what extent do similarity judgments produced by the computational model based on kinematic primitives from the optic flow correlate with human action similarity judgments based on PLDs?
2) To what extent does human accuracy in action-matching tasks relate to similar conditions and results from four different versions of the computational model based on kinematic primitives?
3) To what extent do human judgments of action identification rely on the kinematic features of the actions rather than higher level action semantics?

We addressed these questions using a computational model that derives action primitives based on kinematic features (OF, velocity, acceleration, and change in direction) from biological motion regularities [29]. In Section II, we describe the hand action data set used, followed by a detailed description of the computational model in Section III. The behavior of the model is contrasted with the results of three studies with human participants. Experiment 1 (Section IV) involves an action similarity task (AST) in which participants judge which of two PLD actions is most similar to a PLD target action. The computational model in Section III performs the similarity task by learning to classify actions (using dictionary learning) based on a linear combination of kinematic primitives (sparse coding technique). The similarity values are then used as the matching criterion for the target actions. In Experiment 2 (Section V), participants are given an action matching task (AMT) and we assess to what extent the different model variant representations of actions based on action modality and usage of kinematic features can produce judgments similar to humans. Finally, in Experiment 3 (Section VI), we use an action identification task (AIT) to distinguish between the use of kinematic features and semantic level identification in action similarity judgments made by humans. In other words, are humans relying on high-level semantic features for their similarity judgments rather than on low-level kinematics? This article is concluded with a discussion in Section VII.

TABLE I
ACTIONS WITH THEIR REFERRED TERM

Ref. term	Action
Carrot	Grating a carrot
Cut	Cutting a loaf of bread
Dish	Cleaning a dish
Eat	Eating a slice of bread
Eggs	Beating eggs
Lemon	Squeezing a lemon
Mezzaluna	Using a mezzaluna knife
Mixing	Stirring a mixture,
Openbottle	Opening a bottle
Pan	Pan flip

Ref. term	Action
Pestare	Crushing leaves
Pouring	Pouring water
Reaching	Reaching an object
Rolling	Rolling dough
Salad	Rotating salad chopper
Salt	Using a salt shaker
Spread	Spreading cheese on bread
Table	Cleaning table
Transport	Transporting an object

The key results show that humans and the model were highly accurate in similarity and matching tasks. There was no significant gap between the computational model and human performance. There were, however, noticeable differences in the model variants relying on the velocity profile in terms of selection bias and false hit rates compared to human performance, which will be discussed. There was also a clear result that humans were not able to reliably identify the PLD actions, which strongly suggests that semantic-level processing appears not to contribute to the similarity and matching tasks.

II. HAND ACTION STIMULI

The stimuli used in this study are taken from the multiview cooking actions (MoCAs) data set [30] (available for download from https://github.com/Malga-Vision/MoCA-Project). The full data set includes motion capture data and video recordings of upper body actions executed by one actor in a cooking scenario. All the actions are hand-based and involve some kind of object manipulation. For more details about the actions in the data set (see [30], [31]). This data set was chosen to test action similarity, as hand-based actions cover a wide range of complexity with various movements, and most day-to-day activities involve hand actions. Additionally, since our focus is on kinematic primitives, the several intricacies of manipulative hand movements would cover a variety of kinematic feature interplay. Such rich and complex movements albeit small provide an ideal test bed and learning ground for an investigation into action similarity judgments. Furthermore, an established representation of an action in terms of its kinematic features, however small, can be extrapolated to other large body movements of a similar kinematic nature.

For this study, we chose 19 actions from the data set (see Table I). Most of the actions are carried out by the right hand, whereas some involve both hands (e.g., *Mezzaluna* or *Rolling*).

To investigate the similarity of the action, it was necessary to select a single viewpoint to avoid the excessive duration of the experiment with human participants. Therefore, we chose the frontal viewpoint, which is familiar and natural for interaction, especially during the early stages of child development.

However, the model has been shown to recognize actions from multiple points of view [32], paving the way for future investigations of human perception. See Fig. 1 for an example frame for Eggs and its PLD.

Fig. 1. Left image shows a frame of action eggs from a frontal point of view, and the right image shows its PLD. PLDs correspond to the positions of the markers.

Since this study focuses only on the low-level kinematic features of actions, human participants were shown PLDs that limit their recognition leverage from contextual information. Alternatively, the model was designed to extract only kinematic information directly from the videos (see Section III for details).

III. Computational Model

Two different computational models are used to focus on the bi-modality of the MoCA data set: one based on the biological motion detection model described in [29], that relies on RGB videos to extract OF maps of the apparent motion, and the other model is based on the analysis of motion capture data in [33] that uses histogram representations of the markers' trajectories. The OF-based model takes inspiration from the human ability to distinguish between biological and nonbiological motion, an ability exhibited by newborns in which they orient their attention toward biologically moving stimuli [3]. The model exploits the regularities of human motor movement resulting from the two-thirds power law, a well-known invariant of human movement [34], [35], and has also been implemented on an iCub humanoid robot as proof of applicability [29], [36], [37]. The model based on motion capture data, instead, exploits the sparseness and precision of the markers' trajectory in time to gather information about the spatial evolution of the actions, together with velocity distributions.

A. OF-Based Descriptors

The model for recognizing similarities in actions uses primitives of visual motion to understand actions [32]. The approach is to identify necessary and sufficient action subcomponents and use them as visual primitives to form simple motion representations that can reconstruct a wide range of complex actions. A general breakdown of the model is as follows.

1) The OF from the RGB videos is extracted and thresholded for each time instant. For each point in the obtained maps, the tangential velocity (which is the magnitude of the OF velocities) is computed and such values are averaged over the region. The averaged velocities over time give a compact representation of each video. Velocity sequences over time are segmented into submovements (portions). The submovements are derived automatically with setpoints that correspond to a Start, Stop, and Change in the action dynamics, which are the local minima of the velocity profile [38].
2) The submovements obtained from all the actions (19 hand actions) are treated together and given as input to a clustering of K-means, thereby building a unique dictionary of K atoms. With the dictionary, each submovement of the training set is then reconstructed as an approximation of a linear combination of some of the atoms in the dictionary, using the sparse coding technique, and represented as the sequence of weights used for each atom in the reconstruction. At the end of this procedure, given a video representing a given action, the model can describe each submovement u_i as the feature vector $[u_i^1, u_i^2, \ldots, u_i^K]$, where $u^j{}_i$ is the coefficient/weight assigned to each atom (the jth atom, where $j = 1 \cdots K$). Since the representation is sparse, some of the coefficients are equal to 0, and $K = 15$ is the number of atoms in the dictionary.
3) A classification of the actions (19 hand actions) is performed following a supervised approach. A multiclass classifier is built with a one-versus-all approach, where a binary classifier per class (i.e., per action) is built. Therefore, for each action, a binary classifier is trained to discriminate between the representation of that action and all the rest. See Fig. 2 for an example of how *Eat* contributed to the submovement dictionary and how *Transport* is represented via the dictionary primitives.

B. Motion Capture-Based Descriptors

For the descriptors based on the markers' 3-D information, the model (described in [37]) creates representations of the space–time evolution of action instances combining different joints and their variations over time. We start from Motion Capture data and we build 3D+time equally binned histograms by partitioning the volume of positions and instantaneous velocities (i.e., the displacements between two time-adjacent positions) of actions. Histograms are built using 4 out of the 6 joints available. Fig. 3 shows examples of 3D+time histograms representing the spatial occupation (SP) and velocity components (Vel) for the palm marker. Additionally, a combined descriptor (SP + Vel) is also used, where both the spatial and velocity components are considered for each action.

IV. Experiment 1: Action Similarity Task

Experiment 1 explores the extent to which humans and the OF-based computational model can perform action similarity judgments on the chosen set of hand actions. The purpose of this experiment is to see if the kinematic information in the actions is sufficient for humans (kinematic information from PLDs) and the model (kinematic information from OF) to perform reliable similarity judgments.

A. Human—Action Similarity Task

The human participants performed ASTs on the PLDs of the actions. The PLDs do not provide any contextual information (the tool used or the setting), thereby limiting the contextual semantics associated with the kinematic features.

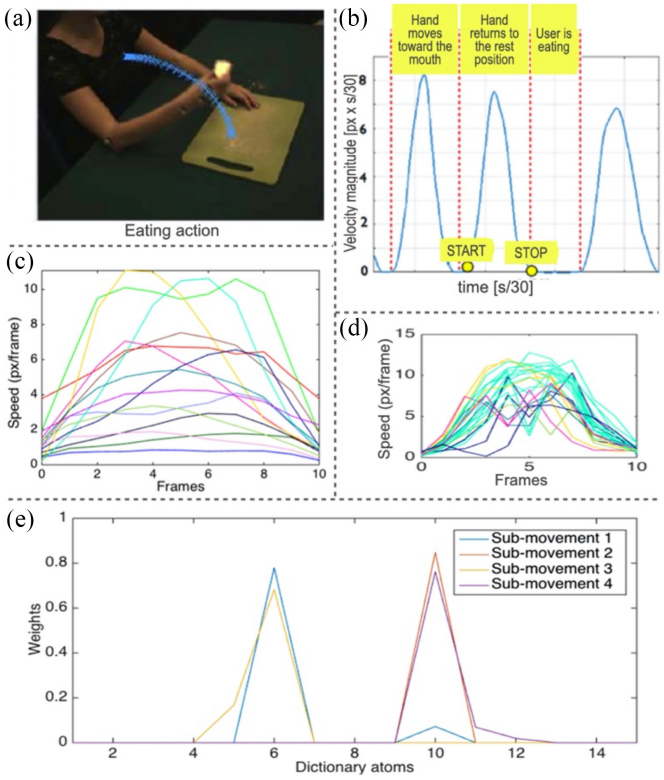

Fig. 2. (a) *Eat* action video from which OF is extracted, (b) identified dynamic instants of Eat action based on set rules and extracted submovements, (c) dictionary of primitives composed of 15 submovements (atoms) extracted from all the 19 actions, (d) submovements extracted from the Eat action, and (e) transport action represented via the dictionary primitives—the submovement 1 has a large contribution from the atom 6, submovement 2 has a large contribution from atom 10, and so on. Images modified from [32] and [38].

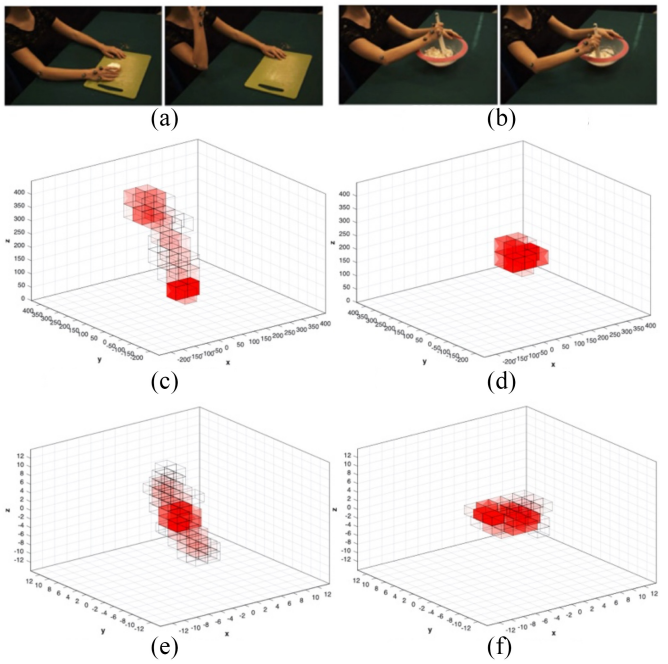

Fig. 3. Example of 3D + t histograms for two different actions. (a) Eating and (b) Mixing. Histogram with SP components for (c) Eating and (d) Mixing. Histogram with Vel components for (e) Eating and (f) Mixing. Images were taken from [30].

1) Participants: Twelve participants (seven males, mean age 27.9 years, age range 24–45 years) from the University of Skövde participated in this experiment. They received information about the task and gave their written informed consent to participate. They received a movie ticket for their participation time. The experiment was carried out in accordance with Swedish law (2003:460) regarding ethical approval and the Declaration of Helsinki of the World Medical Association.

2) Stimuli: The PLDs of the right arm for each of the actions (motion capture data) were generated using Biomotiontoolbox-2 [40] in MATLAB. The PLDs consisted of six dots positioned in the shoulder, elbow, and wrist, and three in the palm region (Fig. 1). Two orientations of the PLDs were used: Upright (UP) and Inverted (INV) (by horizontal flipping of the UP PLD). See Fig. 4 with a trial display of three PLDs (namely, A, B, and T). The stimuli were presented from a frontal point of view (facing the participants) and played at their veridical speed. The experiment was carried out using MATLAB R2014a with Biomotion Toolbox-2 [40] and Psychtoolbox-3 [41]. The stimuli were displayed on a 22-inch HP L2245wg LCD monitor, with a native resolution of 1680 $\times$ 1050 at 60 Hz, a viewable dimension of 29.5 cm $\times$ 47.5 cm (W $\times$ H), and a viewing distance of 100 cm.

3) Procedure: Participants performed a two-alternative forced-choice AST, where they were asked to indicate which of the two stimuli A or B were the most similar to the target stimulus T (see Fig. 4). All possible permutations of the 19 actions were tested. There were two types of trials: 1) one of the alternatives matched the target and 2) none of the alternatives matched the target. Each trial lasted only 4 s, and participants had to respond within the same period. Upon failure to respond in 4 s, the next trial started.

Participants were informed of the corresponding physical characteristics of the PLD, the viewpoint, and the orientations, but no information on the actions themselves was provided, just that they were performing daily actions. The PLDs had random starting frames that played in a continuous loop at 30 FPS. Each response was followed by a fixation cross (0.23°) in the center (500–700 ms). After providing instructions, the participants performed 30 practice tests followed by experiment trials.

The permutation condition (for stimuli A, B, and T) was: a trial may have all three stimuli as different actions ($A \neq B \neq T$) or a trial may have only one other action (A or B) which is the same as the target ($A = T$ OR $B = T$) and $A \neq B$. The total trial number was 6498 and evenly distributed (with balancing) amongst the 12 participants with each participant getting 541 trials (one participant got 547).

B. Model—Action Similarity Task

1) Stimuli: For a given trial, the target action video was fed to the model, and the model extracted OF from the video and computed the motion descriptor only OF was used for each frame as described in Section III. The frontal viewpoint was used for both the training and the model testing.

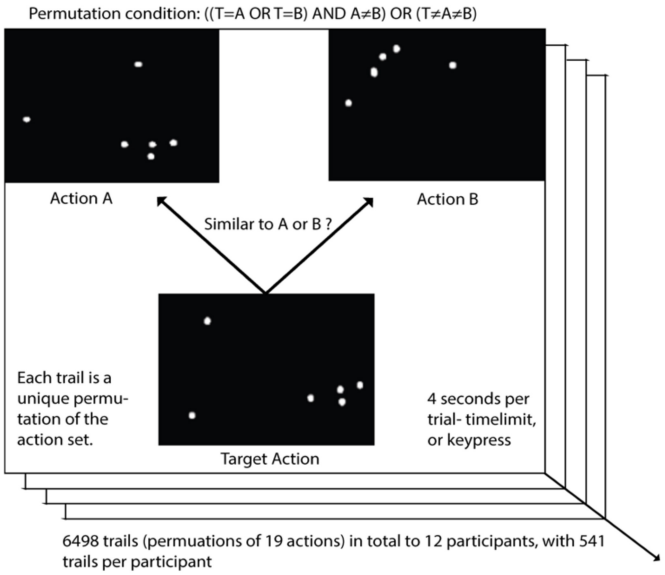

Fig. 4. Experiment 1 (AST) design for both the human participants and model. Note that the model is fed with videos and PLDs for human participants.

2) Procedure: There were 19 action classifiers trained on each of the 19 chosen actions. For classification, a regularized least squares (RLS) classifier was used, which adopted the library GURLS [39] for an efficient implementation of RLS. The radial basis function (RBF) was used as the kernel. The model performed an AMT where it was presented with the target (T) action video and two action classifiers (A and B). These two classifiers competed to see which one of them (A or B) was the most similar to T. So for a given trial (A: *Eat*; B: *Rolling*; and T: *Eggs*), two action classifiers trained on *Eat* (A) and *Rolling* (B) competed, and the classifier with the higher similarity score (toward *Eggs*) won the trial. To simulate the constraints of a viewing period that a human participant would have, random instances of the stimuli were considered, where an instance was one submovement of the action (e.g., in the case of *Mixing*, one half-circular rotation of the palm would be considered one submovement). The similarity measure was computed by averaging the similarities between 10 random instances of the action.

C. Result

Fig. 5 shows the confusion matrix of the responses given by the humans [Fig. 5(a)] and the computational model with OF descriptors [Fig. 5(b)]. The measure reflects the matching rate, with target actions on the y-axis and matched action or the action classifier with a higher score on the x-axis. A cell (i, j) has the match of the ith target action matched with the jth action. Diagonal cells ($i = j$) (accuracy cells) indicate the matching rate when either A or B matched the target action, that is, correctly identified (accuracy measure). Nondiagonal cells ($i \neq j$) report the similarity matching rate when the target action was matched with another action. So, the diagonal cells indicate the accuracy of correctly identifying/matching identical actions, whereas the nondiagonal cells indicate similarity toward other actions. Furthermore, we computed a selection bias (bias of an action in the given choice to get picked), by averaging the measures in column$_j$ minus the respective accuracy cell, giving the mean selection bias (%) of the jth action classifier.

Accuracy measures show that both the human participants ($M = 84\%$, $SD = 8.7$) and OF ($M = 82\%$, $SD = 17.8$) could reliably identify the correct target action (when $A = T$ OR $B = T$), with no significant difference between them. Fig. 5(c) shows a closer comparison of accuracies across the different actions, where we see that except for a few actions (*Lemon, Pestare,* and *pouring*), their performance on most of the actions was at par with each other. There was no significant difference between the human and OF selection-bias measures. The selection bias measures average 40% for both human and OF. This measures the result that is driven primarily by the fact that the experiment design has trials with no correct answer ($A \neq B \neq T$), so the responses end up in the nondiagonal cells which sum up in the selection-bias measure. Nonetheless, if a difference in selection bias between the human and the model shows up, it will be driven by the differences in their accuracies (diagonal cells), which then will spill over to the nondiagonal cells. As the result shows this was not the case here.

D. Discussion

The results demonstrate the extent to which humans and the model with OF descriptor could perform the action similarity judgment when the trials include or did not include matching actions (correct answer). This task reveals two things: 1) the precision of their judgments in trials where there was a correct answer and 2) selection biases in their judgments indicating (possible clusters of) actions that are similar to each other. In terms of precision, both the human (84%) and model (82%) have shown they can reliably identify the same action, but when it comes to selection bias we need further investigation in order to make sense of its distribution, i.e., are they false-hits, or biases.

Furthermore, the nondiagonal cells in conjunction with selection bias do not seem to reveal any strong clusters of similar actions. Nonetheless, this result provides a solid ground to further investigate the role of kinematics in action similarity judgment from a human-centered computational modeling point of view.

This experiment shows that further investigation into action similarity requires a simpler task of judging action similarity, where we can dissect the nondiagonal measures into a clear separation of false-hit and selection bias. Therefore, we switch to a simpler case of an action-matching task where there is always one action alternative that matches the target.

V. Experiment 2: Action-Matching Task

The second experiment extends the study further by 1) modifying the AST to a matching task for all the trials and 2) comparing human performance with the model and following descriptors: velocity from OF, $3D + t$ spatial component (SP), $3D + t$ velocity component (Vel), and, $3D + t$ spatial + velocity component (SP + Vel). The new task will test the

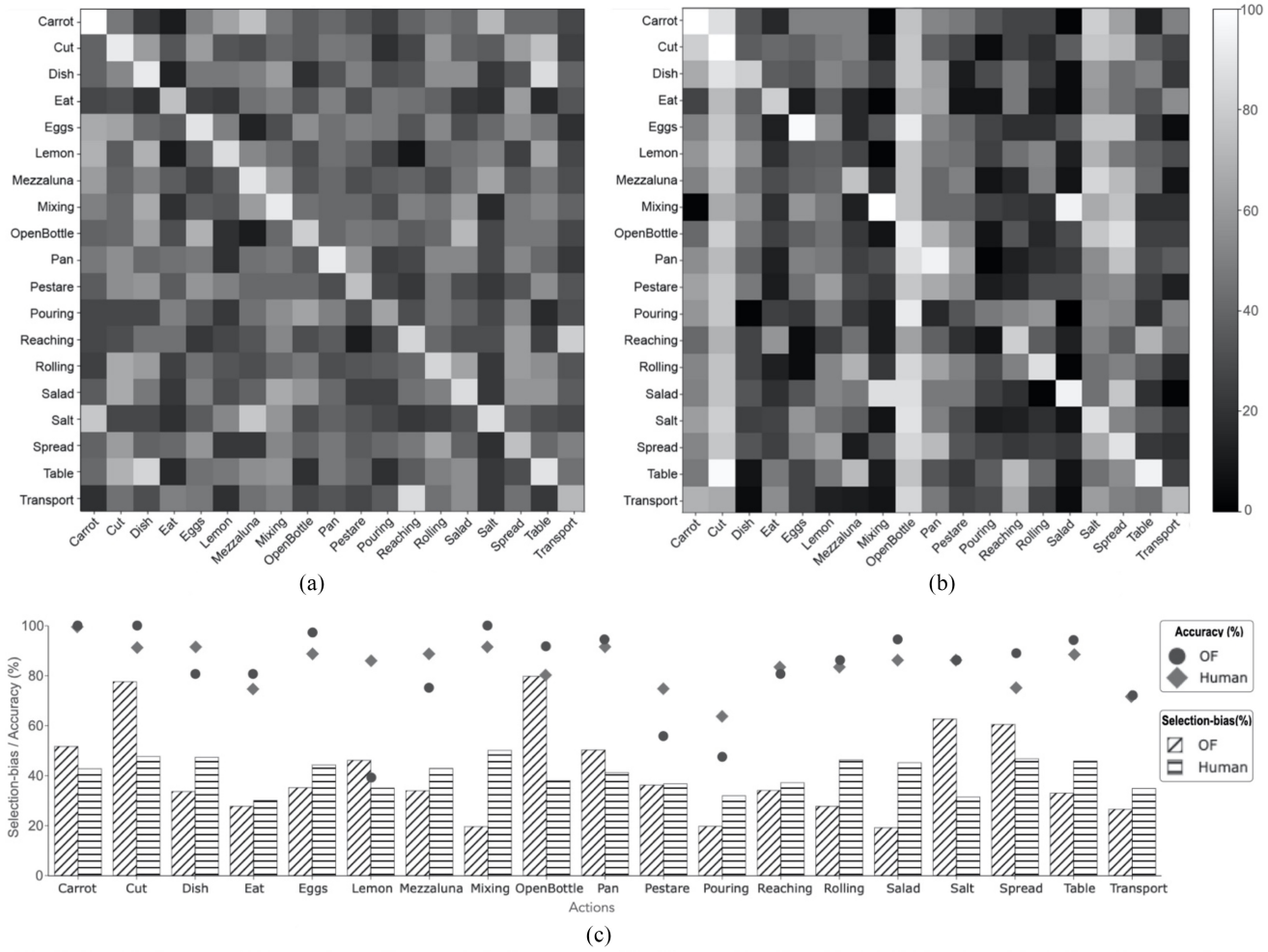

Fig. 5. (a) and (b) Confusion matrices with mean similarity measures (%) of (a) human and (b) model with OF descriptor. (c) Comparison of accuracy (%) and selection-bias(%) of human and model with OF descriptor.

matching capability, i.e., to judge which of the two alternative actions matches the target action based on kinematics, and thereby each trial has a correct answer.

A. Human—Action Matching Task

The stimuli and the procedure were similar to Experiment 1 with a difference in design (see Fig. 6). Each trial consisted of the triad A, B, and T, with the condition ($T = A$ OR $T = B$) AND $A \neq B$, i.e., one of the actions A or B was always the same as the target action (T). With the given condition, the unique permutations for 19 actions give 684 trials. Additional trials were included to assess the implicit semantic access of participants, which tested their performance as a function of orientation: upright (UP) and inverted (INV) PLDs. If participants perform significantly poorly for the INV PLDs in contrast to the UP PLDs (inversion effect), this might indicate implicit semantic access for the UP PLDs.

1) Participants: Twelve subjects (five males, mean age 31.4 years, age range 24–46 years) with normal (or corrected) vision participated. They received information about the task and gave their written informed consent to participate. They received a movie ticket for their participation time. The experiment was carried out in accordance with Swedish law (2003:460) regarding ethical approval and the Declaration of Helsinki of the World Medical Association.

2) Stimuli and Procedure: Same stimuli as in Experiment 1 (human-AST) were used. Participants performed an AMT in which they viewed three actions (A, B, and T) in one frame, and they had to indicate (via keypress) which of the two stimuli A or B was the same as the T stimulus. The rest of the procedure was similar to Experiment 1.

The experiment consisted of three independent variables in a mixed design; orientation (UP/INV, within-subjects), block order (UP-INV/INV-UP, between subjects), and actions (19 actions, random variable). See Fig. 6 for a schematic description. Block orders (UP-INV and INV-UP) were balanced between subjects, with six participants viewing UP-INV. Individual trial orders within blocks were randomized. The total number of trials was 1368×12 (subjects) $= 16\,416$ trials.

B. Model—Action-Matching Task

The model performed an equivalent version of the human—AMT. The stimuli and the procedure were similar to

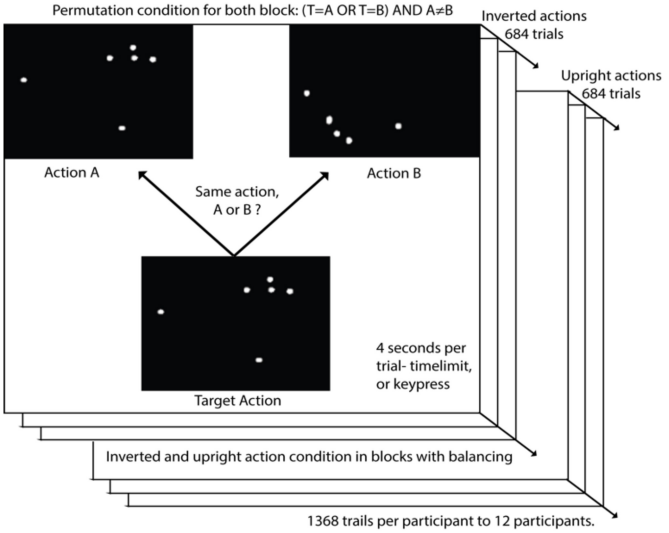

Fig. 6. Experiment-2 (AMT) design for the models and the human participants. Similar to Experiment 1, the models are fed with video, and PLDs for the human participants.

TABLE II
MEANS (%) AND MSE OF ACCURACY (ACC), SELECTION BIAS (SB), AND FALSE-HIT (FH)

	Different Measures in % | *Means, (SD)*			Mean Squared Error		
	Acc	SB	FH	Acc	SB	FH
		Equal Means; Different SD		*(Human as baseline)*		
Human	92.7%, (4.8)	5.7%, (2.4)	(2.9)	—	—	—
OF	85.7%, (16.8)	14.3%, (15.8)	(16.8)	321.5	307.8	347.1
SP	94.3%, (9.3)	2.7%, (1.5)	(5.8)	103.1	14.9	54.9
Vel	89.0%, (14.8)	9.3%, (5.7)	(15.4)	266.5	46.7	270.4
SP+Vel	96.2%, (3.4)	2.1%, (1.3)	(3.2)	30.9	18.3	32.5

Experiment 1 with the experiment design adjusted to the human—AMT. Although for the model there was no INV trial block. The total number of trials conducted was 684 × 24 (times) = 16 416 in randomized order.

C. Results

The results are presented in confusion matrices for humans (H) and the models, with the OF, SP, SP + Vel, and Vel descriptors in Fig. 7. The results are analyzed in terms of the accuracy, false-hit (frequency of an action to get incorrectly picked as the target), and selection-bias.

1) Human Versus Model: Table II shows the comparison of the means for the accuracy (%), selection-bias (%), and false-hit (%) measures in two separate columns—first the actual measures, and then the mean-squared error (MSE) by taking the human results as the baseline. Selection bias and false-hit have the same means (actions collapsed), as they are derived from error (nondiagonal cells), but they vary in their standard deviations.

The average performance in terms of accuracy was high for human participants as well for all variants of the model (c.f. Table II). The MSE for each variant of the model shows how far it is from the human participants' measure (zero = same as human). Considering all measures, the SP+Vel model is the closest to human performance. As can be seen in Fig. 8, the measures vary greatly for individual actions, where

TABLE III
RT AND ACCURACY MEANS, (SD) ACROSS CONDITIONS

Measure	Orientation	Block-Order		
		UP – INV	INV – UP	*average*
RT	**Upright** *(UP)*	1.98, (0.22)	1.67, (0.25)	1.83, (0.27)
(in seconds)	**Inverted** *(INV)*	1.83, (0.21)	1.75, (0.25)	1.79, (0.22)
average		1.91, (0.22)	1.71, (0.24)	
Accuracy	**Upright***(UP)*	92.3, (4.6)	93.5, (1.4)	92.9, (3.3)
(percentage)	**Inverted***(INV)*	92.5, (5.2)	92.3, (1.6)	92.4, (3.7)
average		92.4, (4.6)	92.9, (1.6)	

SP+Vel stands out by providing human-level performance on all actions. Collectively, the selection bias appears to be coming from a few selected action classifiers—*Openbottle, Cut, Spread,* and *Salt*. Similarly, false-hit appears to be for a few actions (target)—*Pestare, Lemon, Pouring*, and *Eggs*.

The selection-bias was lowest for SP+Vel (M = 2.1%, SD = 1.3, MSE = 18.3). Human participants also showed a relatively low selection bias (M = 5.7%, SD = 2.3). While OF showed a comparatively higher selection-bias (M = 14.3%, SD = 15.8, MSE = 307) and a high false-hit (M = 14.3%, SD = 15.4, MSE = 347), closely followed by Vel (M = 9.3%, SD = 15.4, MSE = 2 70).

The measures (accuracy, selection bias, and false hit) were tested with ANOVA, which showed a significant difference between the five conditions (Human, OF, SP, Vel, and SP+Vel), with $F(4, 90) = 2.72$, $p < 0.05$ for accuracy, $F(4, 90) = 9.43$, $p < 0.0001$ for selection-bias, and $F(4, 90) = 6.93$, $p < 0.0001$ for false-hit. A post hoc Tukey HSD test revealed the pairwise difference trend, where there was no significant difference in accuracy between humans and the models. The only significant difference in accuracy was between OF and SP+Vel ($p < 0.05$, HSD [0.05] = 12.25). For selection bias there was a significant difference between OF and others; OF versus Human ($p < 0.05$, HSD [0.05] = 7.54), OF versus other models ($p < 0.01$, HSD [0.01] = 9.08). These pairwise differences seem to be driven by the performance of the few action classifiers and target actions.

2) UP Versus INV—Human: We treated the human performance with respect to the orientation condition (presented only to humans) to see if the participants' performance was not affected by the orientation of the action PLDs.

A 2 orientation (within-subject) × 2 block order (between-subject) mixed ANOVA was performed on the accuracy and RT. Table III presents the RT and accuracy means according to *orientation* × *block order.* The actions were treated as random variables. There was no performance difference between the UP and INV actions. There were no main effects for orientation or block order for both RT and accuracy $p > 0.05$, indicating the lack of an inversion effect. However, there was a significant interaction effect for RT ($F(1,11) = 11.58$, $\eta_p^2 = 0.537$, and $p = 0.007$). The significant difference leading to the interaction effect consisted of faster matching after INV displays (M = 1.982 s, SD = 0.217); $t(10) = 2.31$, $p = 0.043$. Further analyses between UP and INV did not show significant differences, with no simple main effect for

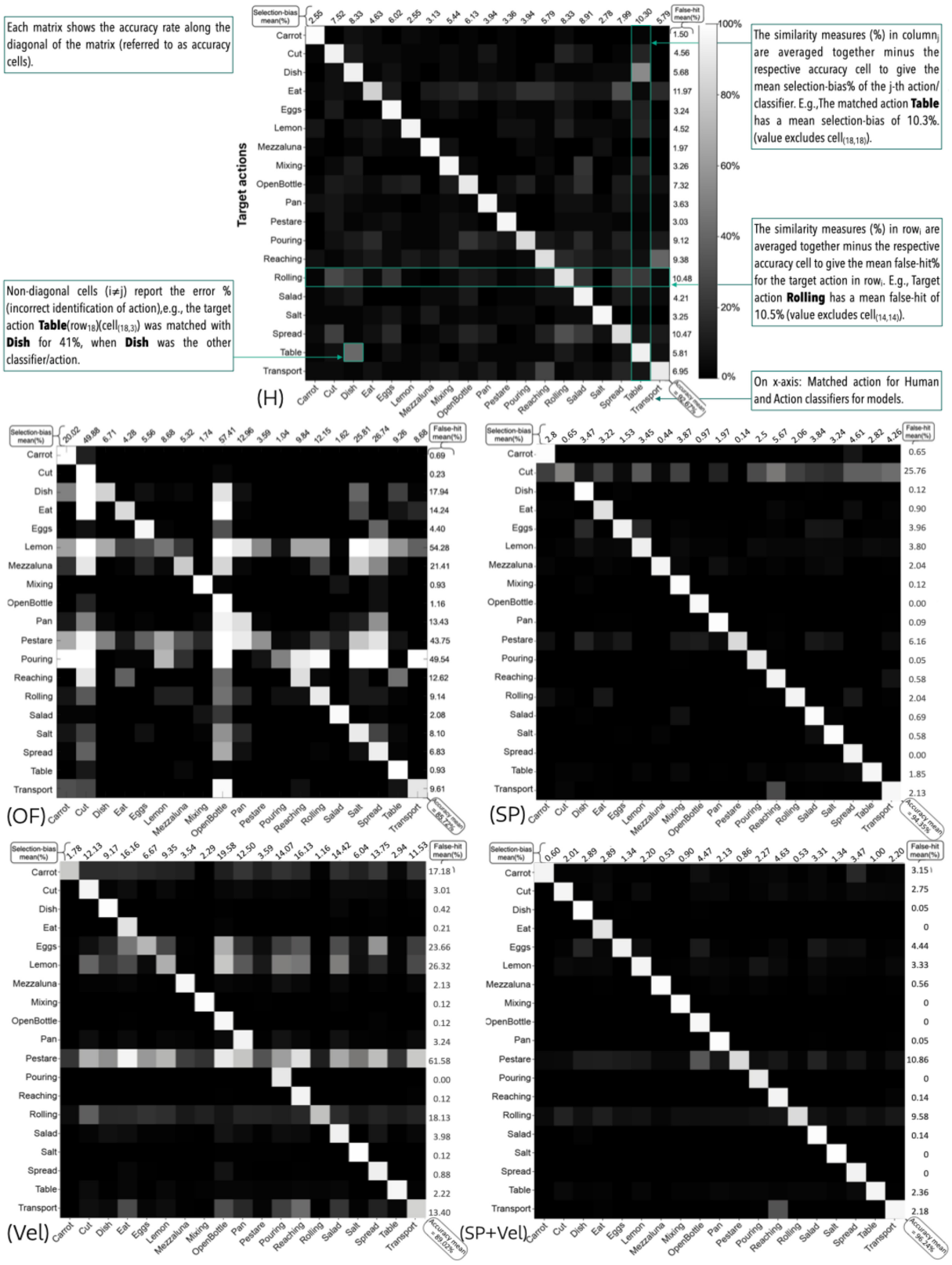

Fig. 7. Confusion matrices with mean similarity measures (%), with target actions (*y*-axis) and matched actions or classifiers (*x*-axis). Matrices show measures for (*H*) human, and the models with the different descriptors OF, SP, Vel, and SP+Vel.

orientation ($p > 0.05$). The interaction effect on RT seems to be driven by the perceptual learning factor in switching from performing on INV PLDs first and then on UP PLDs (fastest RT) which is an easier transition in terms of the perceptual task compared to performing on UP PLDs first and then on INV PLDs.

Fig. 8. Difference measure for accuracy (%), selection bias (%), and false-hit (%) between human (as baseline), and the models with the different descriptors OF, SP, Vel, and SP+Vel.

D. Discussion

1) Accuracy: From the accuracy measure, all the variants of the model and the human participants could reliably identify whether the given actions were the same or not. In taking human performance as a baseline, the SP+Vel performed the closest to humans, closely followed by the SP variant. Their high accuracy reflected the precision of motion capture data, especially when characterizing an action with both spatial occupation and velocity components. The model with Vel and OF descriptors both showed difficulty in matching a few selected actions (e.g., *Lemon, Pestare, and Pouring*). Although Vel and OF relied on different motion data, both relied on the action velocity information. This arises an interesting question about the type of information that could be used for judging action similarity, pointing toward the amount of information necessary to capture the several intricacies of hand action which the human vision system seamlessly captures.

2) Selection Bias: Taking the human selection-bias measure as a baseline, both SP and SP+Vel were at the same level of human performance even with lower selection bias for some actions. Whereas OF had peaked for a few selected actions (*Openbottle, Cut, Carrot, and Salt*), followed by Vel (*Openbottle, Eat, and Pouring*) with a lower difference from human performance. Similar to the observation in accuracy there was a commonality in terms of the kinematic information in use—OF and Vel behaved similarly compared to SP & SP+Vel. Furthermore, the high selection bias came from a small set of action classifiers, indicating a peculiarity with the action such that its velocity components might have affected the selection process.

Going back to the model's dictionary learning process [32], it reveals that these actions have the greatest number of kinematic primitives (atoms) that make up the dictionary primitives (see [32] for detail). So, in other words, these actions contain most of the primitives that make up the submovements of all the 19 actions. Thereby, these actions also correspond to other actions with submovements populated by different atoms. Hence, they have more chances to get confused with other actions, leading to a high selection bias. This also explains why action classifiers with high selection bias also got a high accuracy, as they have sufficient primitives to create a strong representation of their own action.

3) False-Hit: Similar to the case of selection bias, SP+Vel was at the same level of human performance closely followed by SP, where SP had a high false-hit rate for the *cut* action. Whereas *OF* and Vel had a high false-hit rate for a small set of actions (*Lemon, Pestare, and Pouring*). A high false hit shows a lack of descriptive capability, i.e., poor representation of the action by the dictionary primitives. This is in addition to their respective classifiers getting a low selection bias, also pointing toward a lack of sufficient kinematic primitives. False hits for these actions could also result from the classifiers' training process, where necessary and sufficient primitives were not extracted properly for the dictionary.

4) UP Versus INV—Human: The lack of the inversion effect indicates that action semantics did not seem to play much of a role in judging action similarity and that they were mainly relying on low-level kinematic features. This means that there was no implicit access to semantics to aid their judgments. To further ensure that this is mainly due to the fact that humans could not exploit action semantics to aid their judgment, we conducted Experiment 3 to specifically investigate explicit access to semantics.

VI. Experiment 3: Action Identification Task

This experiment addresses the third question of whether human judgments in AST were based solely on the kinematic features of the actions.

A. Human—Action Identification Task

To what extent do humans have access to action semantics that can be used to identify the actions used in the previous two experiments? To this purpose, a five-alternative forced-choice AIT was presented to human participants, where they had to identify the displayed action (in PLD form) from a list of five action labels. Are human judgments in the AST based solely on the kinematic features of the actions rather than higher level action semantics?

1) Participants: Fifty-four Mechanical Turk workers (33 men, mean age 37.33 years, age range 26–73 years) with normal (or corrected) vision and fluency in English participated. They were informed about the task and gave their informed consent to participate. Participants received a monetary compensation of $2.50 for their participation time. The experiment was carried out in accordance with Swedish law (2003:460) regarding ethical approval and the Declaration of Helsinki of the World Medical Association.

2) Stimuli: The trial display consisted of one action PLD at a time followed by five action labels. The PLDs (19 actions) were the same as in Experiment 1: frontal viewpoint played at a veridical speed with UP and INV orientations. The stimuli were displayed using Amazon Mechanical Turk with extensions from psiTurk [42] and jsPsych [43].

3) Procedure: Participants performed an AIT where they were shown an action (target) for 4 s, after which they had to identify (mouse click) the target action label from five action labels (alternatives) within 10 s. The alternatives consisted of the correct label and four randomly chosen labels (from the same pool of 19 action labels) with no repetition. Clicking or failing to respond within 10 s led to the next trial (preceded by a fixation cross for 700 ms). The display orientation (UP or INV) was informed prior to the start. Participants were informed of the PLDs (identical to Experiment 2—human AMT). The instructions were on the screen with example displays. After the instructions, a video of a trial was shown (no practice session). There were questionnaires about the difficulty of the task at the end of the experiment.

The experimental design is identical to Experiment 2—human AMT. The block order (UP-INV and INV-UP) was balanced between the subjects, with 29 participants viewing INV-UP. Individual trial orders were randomized for each participant. The blocks had 19 trials where each trial presented one of the 19 actions; the total number of trials per participant was 38.

B. Results

A selection criterion was used where the mean RT for each participant should exceed 2 s; this was to ensure that the participants diligently performed the task. Therefore, 14 participants were excluded and data from 40 participants were then included in the analysis. Fig. 9 shows the accuracy% (for correct identification) and the selection-bias%. To confirm the reliance of humans on kinematic features for their similarity judgments, we had to rule out explicit semantic level access for the PLDs. If participants perform poorly in identifying the PLDs, regardless of the display orientation, this would strongly suggest a lack of semantic-level access.

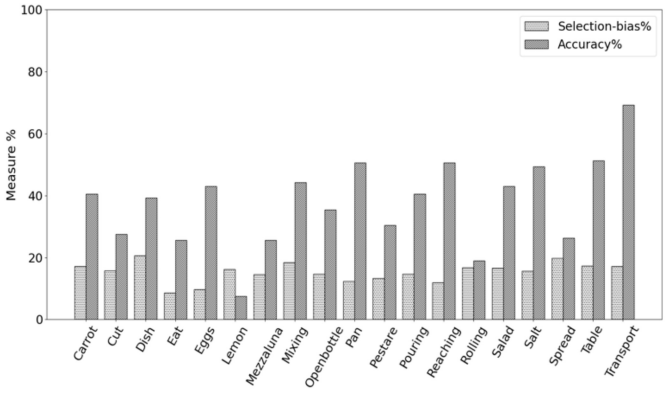

Fig. 9. Accuracy (%) and% selection bias for Experiment 3.

Overall accuracy ($M = 37.85\%$, $SD = 14.17$) indicates poor performance with a mean selection bias of 15.35% ($SD = 3.12$). Participants performed poorly for both UP displays ($M = 38.68\%$, $SD = 15.61$) and INV displays ($M = 35.92\%$, $SD = 16.25$).

A mixed ANOVA of 2 orientation (within-subject) $\times$ 2 block order (between-subject) was performed on the accuracy to check for an inversion effect. The actions were treated as random variables. There was no significant main effect of orientation ($F(1, 39) = 0.966$, $\eta 2p = 0.025$, $p = 0.332$), which shows that there was no performance difference between the action stimuli of UP and INV. The main effect of the block order was also not significant ($F(1, 39) = 1.807$, $\eta 2p = 0.045$, $p = 0.187$). However, there was a significant interaction effect ($F(1, 39) = 6.152$, $\eta 2p = 0.139$, $p = 0.018$). The significant difference leading to the interaction effect consists of a higher accuracy for responses for INV displays ($M = 29.74\%$, $SD = 14.01$) when presented after UP displays ($M = 42.11$, $SD = 16.29$); $t(38) = 2.57$, $p = 0.014$.

C. Discussion

Experiment 3 showed a poor overall accuracy (%), indicating that participants had difficulty identifying the actions of the PLDs displayed. Although most actions were identified above the chance level (i.e., 20%, of 5 options), very few actions had a relatively high accuracy, such as *Transport* = 69%, *Reaching* = 50%, and *Table* = 50%. Despite the poor accuracy, there was no particular selection bias pattern. The kinematic information within the PLDs may not be enough for the participants to recognize the action and choose the correct action labels, which also points to why they did not show any particular selection preference. Observing the results of AIT in light of AMT and AST, no inversion effect was observed for either task, and the poor accuracy in AIT indicates that the participants had very limited access, if any, to semantics in AMT and AST. Hence, we show that humans mainly rely on the kinematic features of the actions to perform AMT and AST—similar to the model.

VII. Conclusion

Three consecutive experiments compared human performance on various action recognition tasks with the performance of a computational model (see Section III for details). The results from Experiment 1 showed that the

model was close to human performance on a fairly difficult similarity judgment task, despite the fact that very limited information in the form of average OF was used. Experiment 2 provided further insight by showing that the majority of actions could be matched based on limited information within a single feature domain (velocity or spatial), while both velocity and spatial information were necessary to reach human performance on all actions. Experiment 3 indicated that human participants lacked access to semantic information in their judgments, further strengthening the conclusion that velocity and spatial features are both used in HAR, but that most actions can be recognized based on a single feature domain if necessary. Further research is needed to understand how humans utilize the velocity and spatial features to judge action similarities—do they form kinematic primitives similar to the model, and if so, how?

The current work provides insight into the potential mechanisms supporting action similarity detection in humans, providing a pathway toward implementing similar models in machines. The approach has a developmental inspiration, in that it is based on an existing model of the ability of newborns (biological motion detection [37]) to assess how far this simple representation allows one to go in terms of a novel and more complex skill, such as the detection of similarity of actions. It is important to note that progressive development could continue with more complex social competencies. In fact, for humans, the detection of action similarity plays a fundamental role in imitation. In particular, according to the similarity model [44], kinematic similarity increases the predictability of the action. Imitation, in turn, supports the development of action understanding. For example, several researchers have suggested that the experience of being imitated is crucial in the development of the mirror neurons system (e.g., [45] and [46]). In this context, the child's ability to judge the kinematic similarity between her and her caregiver's actions would support the child's ability to mimic, which is another step toward understanding the action.

In a similar vein, the topic of imitation has also been widely investigated in robotics (e.g., [47], [48], [49], and [50]) and has important implications for the domain of learning from demonstration [51]. Additionally, for this application, the possibility of detecting similarities in actions and performing actions that closely resemble that of the human partner could increase the intuitiveness and efficacy of the interaction.

An aspect worthy of further attention refers to the performance based on the type of action in use. Humans viewed PLDs to achieve a high level of performance, similar was the case for the model variant relying on the same PLDs, whereas the OF variant relied on the video feed and also performed reliably high with comparatively lower performance (in terms of selection-bias and false-hit). The experiments presented here showcase that an efficient and effective strategy of utilizing kinematic primitives of relevant motion is possible in both modalities. Additionally, such a strategy could serve as a part of a more complex action-recognition model mimicking the robustness of human action judgment capability.

This work uses kinematics in the two different modalities highlighting the importance of understanding the significance of biological motion parameters and how they can be utilized toward action understanding both from the point of view of Human cognition and HAR models. Many recent approaches to action recognition in computer vision rely on deep neural networks, e.g., [25] and [52] (see the detailed review of approaches in [53]). These are quite different from the presented model in that they use many more parameters in their descriptors, providing much more powerful classifiers. As a result, deep networks are often described as black-box models, providing powerful input–output mappings but providing little information on the internal mechanisms necessary to identify the stimuli. With the work presented here, we argue for the need to investigate human action understanding to enable more white-box methods toward designing systems such that they understand actions with the human at the center, i.e., a step closer to human-centered computational modeling.

ACKNOWLEDGMENT

This work has been partially carried out at the Machine Learning Genoa (MaLGa) Center, Università di Genova, Italy.

REFERENCES

[1] B. Tversky, *Mind in Motion: How Action Shapes Thought*. London, U.K.: Hachette U.K., 2019.
[2] G. Yovel and A. J. O'Toole, "Recognizing people in motion," *Trends Cogn. Sci.*, vol. 20, no. 5, pp. 383–395, 2016.
[3] F. Simion, L. Regolin, and H. Bulf, "A predisposition for biological motion in the newborn baby," *Proc. Nat. Acad. Sci.*, vol. 105, no. 2, pp. 809–813, 2008.
[4] J. C. Stapel, "The development of action perception," in *Modelling Human Motion*. Cham, Switzerland: Springer, 2020, pp. 73–101.
[5] R. Sifre, L. Olson, S. Gillespie, A. Klin, W. Jones, and S. Shultz, "A longitudinal investigation of preferential attention to biological motion in 2-to 24-month-old infants," *Sci. Rep.*, vol. 8, no. 1, pp. 1–10, 2018.
[6] P. E. Hemeren, *Mind in Action*, Lund Univ. Cogn. Stud., Lund, Sweden, 2008.
[7] C. E. Watson and L. J. Buxbaum, "Uncovering the architecture of action semantics," *J. Exp. Psychol. Human Percept. Perform.*, vol. 40, no. 5, p. 1832, 2014.
[8] M. Giese and M. Lappe, "Measurement of generalization fields for the recognition of biological motion," *Vis. Res.*, vol. 42, no. 15, pp. 1847–1858, 2002.
[9] C. Catmur and C. Heyes, "Is it what you do, or when you do it? The roles of contingency and similarity in prosocial effects of imitation," *Cogn. Sci.*, vol. 37, no. 8, pp. 1541–1552, 2013.
[10] Z. Sun, Q. Ke, H. Rahmani, M. Bennamoun, G. Wang, and J. Liu, "Human action recognition from various data modalities: A review," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 3, pp. 3200–3225, Mar. 2023, doi: 10.1109/TPAMI.2022.3183112.
[11] G. Qing and H. Hui, "Standardized judgment method of shooting training action based on digital video technology," *Sci. Program.*, vol. 2021, Dec. 2021, Art. no. 4725875.
[12] J. Li, H. Cui, T. Guo, Q. Hu, and Y. Shen, "Efficient fitness action analysis based on spatio-temporal feature encoding," in *Proc. IEEE Int. Conf. Multimedia Expo Workshops (ICMEW)*, 2020, pp. 1–6.
[13] S. C. Akkaladevi and C. Heindl, "Action recognition for human robot interaction in industrial applications," in *Proc. IEEE Int. Conf. Comput. Graph. Vis. Inf. Security (CGVIS)*, 2015, pp. 94–99.
[14] I. Rodomagoulakis et al., "Multimodal human action recognition in assistive human–robot interaction," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP)*, 2016, pp. 2702–2706.
[15] J. Hayakawa and B. Dariush, "Recognition and 3D localization of pedestrian actions from monocular video," in *Proc. IEEE 23rd Int. Conf. Intell. Transp. Syst. (ITSC)*, Sep. 2020, pp. 1–7.
[16] M. Yu, J. Xu, J. Zhong, W. Liu, and W. Cheng, "Behavior detection and analysis for learning process in classroom environment," in *Proc. IEEE Front. Educ. Conf. (FIE)*, 2017, pp. 1–4.
[17] J. Yin, J. Han, C. Wang, B. Zhang, and X. Zeng, "A skeleton-based action recognition system for medical condition detection," in *Proc. IEEE Biomed. Circuits Syst. Conf. (BioCAS)*, 2019, pp. 1–4.

[18] L. G. Clift, J. Lepley, H. Hagras, and A. F. Clark, "Autonomous computational intelligence-based behaviour recognition in security and surveillance," in *Proc. Counterterrorism, Crime Fighting, Forensics, Surveillance Technol. II*, 2018, pp. 173–179.

[19] J. Qin, L. Liu, Z. Zhang, Y. Wang, and L. Shao, "Compressive sequential learning for action similarity labeling," *IEEE Trans. Image Process.*, vol. 25, pp. 756–769, 2015.

[20] O. Kliper-Gross, T. Hassner, and L. Wolf, "The action similarity labeling challenge," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 3, pp. 615–621, Mar. 2012.

[21] P. E. Hemeren and S. Thill, "Deriving motor primitives through action segmentation," *Front. Psychol.*, vol. 1, p. 243, Jan. 2011.

[22] D. Kulić, D. Kragic, and V. Krüger, "Learning action primitives," in *Visual Analysis of Humans*. London, U.K.: Springer, 2011, pp. 333–353.

[23] Y. Kong and Y. Fu, "Human action recognition and prediction: A survey," 2018, *arXiv:1806.11230*.

[24] A. Sarkar, A. Banerjee, P. K. Singh, and R. Sarkar, "3D human action recognition: Through the eyes of researchers," *Expert Syst. Appl.*, vol. 193, May 2022, Art. no. 116424.

[25] Z. Tu et al., "Multi-stream CNN: Learning representations based on human-related regions for action recognition," *Pattern Recognit.*, vol. 79, pp. 32–43, Jul. 2018.

[26] H. Wang and C. Schmid, "Action recognition with improved trajectories," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2013, pp. 3551–3558.

[27] M. Barnachon, S. Bouakaz, B. Boufama, and E. Guillou, "Ongoing human action recognition with motion capture," *Pattern Recognit.*, vol. 47, no. 1, pp. 238–247, 2014.

[28] S. Vantigodi and V. B. Radhakrishnan, "Action recognition from motion capture data using meta-cognitive RBF network classifier," in *Proc. IEEE 9th Int. Conf. Intell. Sens. Sens. Netw. Inf. Process. (ISSNIP)*, 2014, pp. 1–6.

[29] A. Vignolo, N. Noceti, A. Sciutti, F. Rea, F. Odone, and G. Sandini, "The complexity of biological motion," in *Proc. Joint IEEE Int. Conf. Develop. Learn. Epigenet. Robot. (ICDL-EpiRob)*, 2016, pp. 66–71.

[30] E. Nicora, G. Goyal, N. Noceti, A. Vignolo, A. Sciutti, and F. Odone, "The MoCA dataset, kinematic and multi-view visual streams of fine-grained cooking actions," *Sci. Data*, vol. 7, no. 1, pp. 1–15, 2020.

[31] D. Malafronte, G. Goyal, A. Vignolo, F. Odone, and N. Noceti, "Investigating the use of space-time primitives to understand human movements," in *Proc. Int. Conf. Image Anal. Process.*, 2017, pp. 40–50.

[32] A. Vignolo, N. Noceti, A. Sciutti, F. Odone, and G. Sandini, "Learning dictionaries of kinematic primitives for action classification," in *Proc. Int. Conf. Pattern Recognit. (ICPR)*, 2020, pp. 5965–5972.

[33] E. Nicora, G. Goyal, N. Noceti, and F. Odone, "The effects of data sources: A baseline evaluation of the MoCA dataset," in *Proc. Int. Conf. Image Anal. Process.*, 2019, pp. 544–555.

[34] P. Viviani and N. Stucchi, "Biological movements look uniform: Evidence of motor-perceptual interactions," *J. Exp. Psychol. Human Percept. Perform.*, vol. 18, no. 3, p. 603, 1992.

[35] M. J. E. Richardson and T. Flash, "Comparing smooth arm movements with the two-thirds power law and the related segmented control hypothesis," *J. Neurosci.*, vol. 22, no. 18, pp. 8201–8211, 2002.

[36] A. Vignolo, F. Rea, N. Noceti, A. Sciutti, F. Odone, and G. Sandini, "Biological movement detector enhances attentive skills of humanoid robot iCub," in *Proc. IEEE-RAS Int. Conf. Humanoid Robots*, 2016, pp. 338–344.

[37] A. Vignolo, N. Noceti, F. Rea, A. Sciutti, F. Odone, and G. Sandini, "Detecting biological motion for human–robot interaction: A link between perception and action," *Front. Robot. AI*, vol. 4, p. 14, Jun. 2017.

[38] F. Rea, A. Vignolo, A. Sciutti, and N. Noceti, "Human motion understanding for selecting action timing in collaborative human–robot interaction," *Front. Robot. AI*, vol. 6, p. 58, Jul. 2019.

[39] A. Tacchetti, P. S. Mallapragada, M. Santoro, and L. Rosasco, "Gurls: A toolbox for regularized least squares learning," Comput. Sci. Artif. Intell. Lab., Massachusetts Inst. Technol., Cambridge, MA, USA, Rep. MIT-CSAIL-TR-2012-003, 2012.

[40] J. J. van Boxtel and H. Lu, "A biological motion toolbox for reading, displaying, and manipulating motion capture data in research settings," *J. Vis.*, vol. 13, no. 12, p. 7, 2013.

[41] M. Kleiner, D. Brainard, and D. Pelli, "What's new inpsychtoolbox-3?" *Perception*, vol. 36, no. 14, pp. 1–16, 2007.

[42] T. M. Gureckis et al., "psiTurk: An open-source framework for conducting replicable behavioral experiments online," *Behav. Res. Methods*, vol. 48, no. 3, pp. 829–842, 2016.

[43] J. R. De Leeuw, "jsPsych: A Javascript library for creating behavioral experiments in a Web browser," *Behav. Res. Methods*, vol. 47, no. 1, pp. 1–12, 2015.

[44] J. Hale and A. F. D. C. Hamilton, "Cognitive mechanisms for responding to mimicry from others," *Neurosci. Biobehav. Rev.*, vol. 63, pp. 106–123, Apr. 2016.

[45] C. Catmur, V. Walsh, and C. Heyes, "Associative sequence learning: The role of experience in the development of imitation and the mirror system," *Philos. Trans. Royal Soc. B*, vol. 364, no. 1528, pp. 2369–2380, 2009.

[46] S. S. Jones, "Infants learn to imitate by being imitated," in *Proc. 10th Int. Conf. Devel. Learn.*, 2006, pp. 1–6.

[47] K. Yasuo, Y. Yasuaki, I. Masayuki, and I. Hirochika, "From visuomotor self-learning to early imitation—A neural architecture for humanoid learning," in *Proc. IEEE Int. Conf. Robot. Autom.*, vol. 3, 2003, pp. 3132–3139.

[48] Y. Nagai, Y. Kawai, and M. Asada, "Emergence of mirror neuron system: Immature vision leads to self-other correspondence," in *Proc. IEEE Int. Conf. Develop. Learn. (ICDL)*, vol. 2, 2011, pp. 1–6.

[49] E. A. Billing, T. Hellström, and L. E. Janlert, "Predictive learning from demonstration," in *Proc. 2nd Int. Conf. Agents Artif. Intell.*, 2011, pp. 186–200.

[50] E. A. Billing, T. Hellström, and L.-E. Janlert, "Behavior recognition for learning from demonstration," in *Proc. IEEE Int. Conf. Robot. Autom.*, 2010, pp. 866–872.

[51] B. Argall, S. Chernova, M. Veloso, and B. Browning, "A survey of robot learning from demonstration," *Robot. Auton. Syst.*, vol. 57, no. 5, pp. 469–483, 2009.

[52] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri, "Learning spatiotemporal features with 3D convolutional networks," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2015, pp. 4489–4497.

[53] D. Wu, N. Sharma, and M. Blumenstein, "Recent advances in video-based human action recognition using deep learning: A review," in *Proc. Int. Joint Conf. Neural Netw. (IJCNN)*, 2017, pp. 2865–2872, doi: 10.1109/IJCNN.2017.7966210.