Computational Visual Media
DOI 10.1007/s41095-xxx-xxxx-x Vol. x, No. x, month year, xx–xx

Article Type (Research/Review)

Object properties inferring from and transfer for human interaction motions

Qian Zheng[1], Weikai Wu[1], Hanting Pan[1], Niloy Mitra[2], Daniel Cohen-Or[3], and Hui Huang[1] (✉)

Abstract Humans regularly interact with their surrounding objects. Such interactions often result in strongly correlated motion between humans and the interacting objects. We thus ask:"Is it possible to infer object properties from skeletal motion alone, even without seeing the interacting object itself?" In this paper, we present a fine-grained action recognition method that learns to *infer* such latent object properties from human interaction motion alone. This inference allows us to *disentangle* the motion from the object property and *transfer* object properties to a given motion. We collected a large number of videos and 3D skeletal motions of the performing actors using an inertial motion capture device. We analyze similar actions and learn subtle differences among them to reveal latent properties of the interacting objects. In particular, we learn to identify the interacting object, by estimating its weight, or its fragility or delicacy. Our results clearly demonstrate that the interaction motions and interacting objects are highly correlated and indeed relative object latent properties can be inferred from the 3D skeleton sequences alone, leading to new synthesis possibilities for human interaction motions. Dataset is available at `http://vcc.szu.edu.cn/research/2020/IT`.

Keywords human interaction motion, object property inference, motion analysis, motion synthesis.

1 Introduction

Digitizing and understanding our physical world are important goals of both computer graphics and computer vision. In natural environments, humans regularly interact with their surrounding objects and, as an effect, such interactions result in strongly correlated motion between humans and the interacting objects. Researchers in experimental psychology show that observers not only can recognize motion categories, but *also infer object properties* by observing corresponding human motion alone, even without directly seeing the object itself [4]. For example, we humans, regularly estimate object properties like the weight, fragility, path width, or shape, by observing either the real action of a human or even a pantomimed or virtual avatar action [39, 40, 50].

One way to computationally exploit such correlated human-object motions under interactions would be to learn object properties by learning correlation with human skeletal motion over time. However, the available datasets for human activity recognition [31, 44] are RGB-D videos, which in general contain significant occlusions that hamper the extraction of unseen acting skeletons. While these videos can be used to broadly classify different actions [34], we still lack suitable datasets specifically designed for inferring fine-scale variations of object properties. Unlike previous efforts on action recognition, we analyze *similar actions* and hence have to learn subtle differences among the same type of the action that reveal latent properties of interacting objects. Inspired by previous works on motion style transfer, which transform an input motion into a new style while keeping its content, we use these latent properties to edit a given motion. For example, given the skeletal motion of a person walking on a wide path, we would like to synthesize the person's skeletal motion when walking on a narrow path.

In our work, we focus on eight typical types of human-object interaction, including lifting a box, moving a bowl, and walking on a path. We collected video and 3D skeletal

1 Shenzhen University, Shenzhen, China. E-mail: qianzheng85@gmail.com, wuweikai0617pk@gmail.com, panhanting95@gmail.com, hhzhiyan@gmail.com(✉)

2 University College London, London, UK. E-mail: n.mitra@cs.ucl.ac.uk.

3 Tel Aviv University, Tel-Aviv, Israel. E-mail: cohenor@gmail.com.

Manuscript received: 2021-1-18; accepted: xxxx-xx-xx

motions of the performing actors using an inertial motion capture device, which do not suffer from occlusions that are unavoidable from video-based recordings. For these interactions, we learn to infer latent properties of the interacting object from the 3D skeleton sequences alone. In particular, we learn to identify the interacting object, by estimating its *property value*, i.e., a particular value of a property, such as 0kg/15kg/25kg for box weight, or empty/full for bowl fragility.

For the inference task, we treat objects' latent property estimation as a fine-grained classification problem by analyzing similar input skeletal motions. Although some properties (*e.g.* the weight) may vary continuously, treating it as a regression problem requires more training samples. We represent a skeleton sequence as a time sequence of graph structure, which encodes the position and speed information of all joints with temporal dynamics. After analyzing per-joint features, we feed it into a recurrent network to recognize the latent object properties. The results obtained demonstrate that the interaction motions and interacting objects are highly correlated, where object property values can indeed be inferred, to a certain accuracy, by just observing human movements. We will show that, comparing with existing works for action recognition, our method achieves higher inference accuracy.

For the synthesis task, we develop a network architecture to disentangle object property from the abstract motion, which allows to create novel skeletal motions by mixing new object properties on target skeletons. We train a deep neural network with a simple encoder-decoder structure to conduct the disentanglement, *i.e.*, the latent space encodes the motion content *without* object property. A motion can then be synthesized given a specific property value.

In summary, we claim the following contributions:

- Learning subtle differences among the same type of motions of humans interacting with an object;
- A property and motion disentanglement network that allows motion synthesis conditioned on target interactions;
- Introducing an extensive interaction dataset for object property inference from motions with 4k+ samples collected from 100 participants, including eight daily interactions (*i.e.*, lifting a box, moving a bowl, walking, fishing, pouring liquid, bending, sitting, and drinking), which will be released.

2 Related Work

Our work analyzes human interaction motion to detect object properties. Therefore, we briefly describe previous approaches that exploit human-object interactions from visual inputs, with a focus on object property inference. Since we use skeleton sequences to represent motions, we also review those related works on skeleton-based action recognition.

Human-object interaction. Human-object interaction detection itself is an important scientific problem [58] with wide practical uses. Recent methods can successfully detect <human, verb, object> triplets from visual inputs [11, 26].

A variety of techniques in shape analysis have been developed to extract functional information of objects and scenes using human-object interaction as cues. An appropriate human pose or action map can be created from an input object [12, 20, 28] or scene [30, 42]; see a survey [19] for more information. The hidden human context was used as a cue for labeling and arranging the scenes [22, 23]. However, there is no work yet solving this inverse problem: *inferring object properties from human motions and/or interactions alone.*

The spatial relationship between the characters and objects in the environment captures the semantics of interactions. Ho *et al.* [16] introduced interaction mesh structure to explicitly represent the spatial relationship for motion retargeting. Later this representation was used for motion comparison [45].

Object property inference. Researchers in psychology reported that observers can make fine discrimination when presented with human motions in visual form. The weight of a box can be *seen* by observing another person lifting and carrying it [40], and the elasticity of a supporting surface can be judged by observing a person walking on that surface [47]. Vaina *et al.* [50] demonstrated that the weight of an object was robustly estimated, while size and shape were harder to estimate by observers. Recently, Podda *et al.* [39] showed that participants were able to identify the weight of the to-be-grasped object from both occluded real and pantomimed movements, solely using available kinematic information. Observers seem to focus most on the duration of the lifting movement to perceptually judge the weight [10]. Some findings suggest observers may integrate multiple sources for object property inference; for example, shape, motion, and optical cues are used when inferring stiffness [43]. Still, we focus on inference from motions alone in this work.

The object classes and their 3D locations can be recovered from motion by exploiting the human-object spatial relations, used for synthetic scene reconstruction [25] and scene arrangement recovery [36]. There is not much effort made to automatically infer other properties. Davis and Gao [9] presented a computational framework that can label the effort of an action corresponding to the perceived

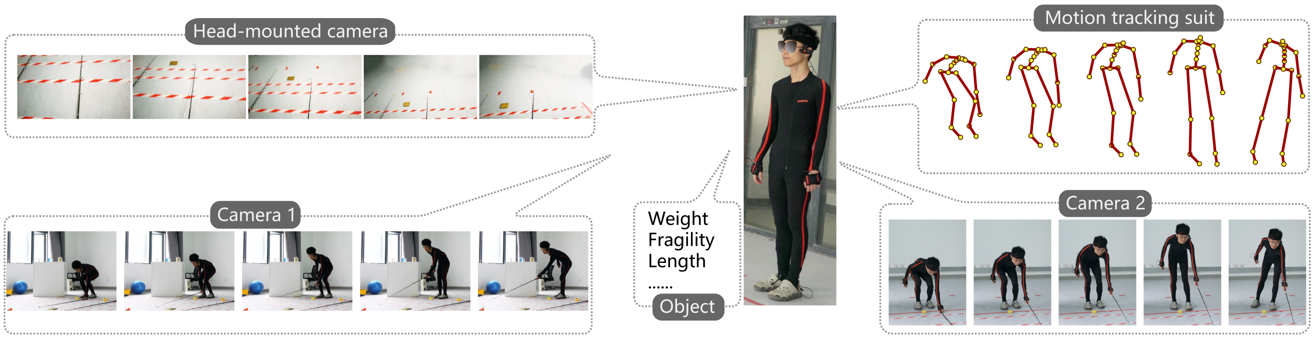


Fig. 1 For each sample, we capture a 3D skeleton sequence by an inertial motion tracking suit, an ego-centric video by a head-mounted camera, two other videos by two cameras placed outside, and the object's geometry along with its properties.

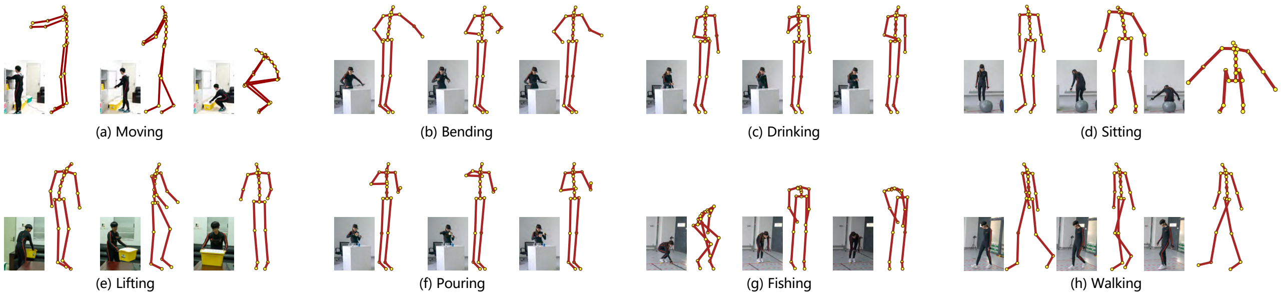


Fig. 2 Eight interaction motions represented in our dataset, which comprises of 4k+ interaction captures across 100 different participants.

level of exertion by the performer. Gupta and Davis [14] did a classification of *heavy/light* objects based on the velocity of ballistic motions detected from video. Integrating a 3D physics engine is another way to infer physical properties, including mass, position, 3D shape, and friction etc., from real-world videos [54, 55].

Action recognition and motion style transfer. With the availability of large-scale skeleton datasets, deep learning is popular for action recognition. Skeleton sequences are indeed the time series of joint positions. The recurrent neural networks, designed to model long-term temporal dependency problems, have been well exploited for skeleton sequences [32, 33, 46]. Skeleton is also a special graph structure representation, and thus graph convolution networks are utilized as well for action recognition [57].

CNN models are able to extract high-level information and have also been used to deal with skeleton sequences. A skeleton sequence can be converted into an image or a 3D tensor, and then fed into a CNN to recognize the underlying action. These methods vary most in the representations of skeleton sequences and network structures. Ke *et al.* [27] represented a skeleton sequence as several images to encode different spatial relationship in-between joints, and then applied pre-trained VGG to extract the features. Li *et al.* [29] represented a skeleton sequence as a 3D tensor, and modeled the global co-occurrence patterns with CNN. Most recently, Aristidou *et al.* [3] used a triplet loss network to map short motion clips to an embedding space, where the distances represent similarity between motion clips. We also utilize graph convolution and RNN to learn object properties from skeletal motions. Nonetheless, we propose to use sub-categorical properties to effectively distinguish fine-grained differences between the motions of the same class.

Another related topic is motion style, which usually represents the mood or identity of a particular character's motion. By analyzing differences between performances of the same content in different styles, researchers have proposed the methods to transform an input motion data into new styles [18, 56, 59]. The object properties and actions are significantly correlated. A particular object property can be only observable in a particular action type, which makes the existing motion style transfer techniques not suitable for our synthesis task.

3 Interaction Motion Dataset Collection

Traditionally, human motion is captured using optical marker-based systems while the markers are placed on the performer. With recent success of deep learning, 2D poses [5, 21, 37, 41, 53] and 3D poses [2, 24, 35, 38, 48, 49] can be extracted directly from RGB or RGB-D video sequences. Large-scale skeletal motion datasets, such as CMU [8], NTU RGB+D [44] and PKU-MMD [31]), are available and driving forces for motion recognition, retrieval and synthesis. However, although these datasets contain human-object interaction motions, the object information are usually unlabeled, and the (partial) joint trajectories are not sufficient to reliably infer 3D object properties. For example, some limbs are very likely to be occluded by the interacting objects. Such occlusions make it very difficult to robustly extract high-quality skeletal motions from monocular or RGB-D videos, even with state-of-the-art pose detection methods. This is particularly true in our setting where we seek subtle motion differences. Therefore, we use inertial measurement units (IMUs) to get 3D human motions that are totally free of occlusions.

Data modalities. We utilize multiple data modalities to construct our dataset. When performing the actions, each subject wore an Xsens MVN inertial motion tracking suit to capture the high-quality 3D skeleton information at 240 frames per second. Each subject was also required to wear a head-mounted camera to capture ego-centric video. Further, we used three uncalibrated cameras to record the subject from three different views, storing three videos at 50 frames per second. For each interacting object, in addition to measuring its size and weight, we also scanned its geometry shape. Fig. 1 presents our capturing scenario and the data modalities of each motion sample collected. Although in this work we only use 3D skeletal information to infer the object properties, we believe that these data modalities are useful for the future research.

Subjects and object interactions. We carefully selected human-object interactions to depict the correlation between human motions and properties of objects. For a good candidate, object property values could be inferred easily from the whole interaction motion alone, but difficultly from a single static frame. Following this rule, we chose eight daily interaction: Walking for estimating *the width* of the path, Fishing for *the length* of a fishing rod, Pouring for *the type* of liquid, Bending for *the stiffness* of a power twister, Sitting for estimating *the softness* of a chair being sat on, Drinking for estimating *the amount* of water inside a cup, LiftingBox for *the weight* of an object be lifted, and MovingBowl for *the fragility* of an object. These motions are shown in Fig. 2. We have invited 100 different subjects for our data collection. They vary in age (20–35), gender (M or F), height (150–195cm) and strength (weak–strong).

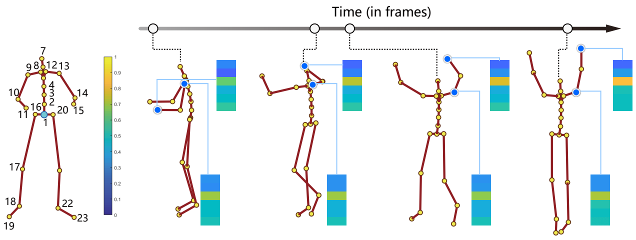


Fig. 3 We represent a skeleton sequence as a tree sequence. The input feature of each joint is represented by its xyz location and velocity in a local body frame coordinate. The cyan point indicates the root (pelvis) of the tree. Each block indicates the joint's feature at a frame.

Here we briefly describe the setting of Walking. Please refer to the appendix for the settings of other interactions.

WALKING. Each subject was asked to walk back and forth on three straight paths of different widths. We simulated the width of a path using line markers to indicate path borders, and asked the subjects do not cross the borders. So we have a total of $3 \times 2 \times 100 = 600$ motion samples.

4 Object Property Inference

4.1 Skeleton sequence representation

The input skeleton data is a sequence of multi-frame tree structure with 3D joints as nodes that form an *action*. As shown in Fig. 3, a skeleton sequence is denoted as a 3D tensor of size $T \times J \times D$, with T representing the frame length, $J = 23$ the total number of joints, and D the feature dimension of each joint, respectively.

Representing a skeleton sequence by joints in xyz locations is common [27, 29, 44]. Some researchers also represent the joints in 3D angles [3]. In our case, the object properties that we aim to estimate are highly correlated with the dynamic properties of motions. As we show in results, joint trajectories (position and velocity representations) can overall help with object property inference.

Each joint is represented by the x, y, and z coordinate in a local body coordinate system with its origin on the pelvis joint (indicated with a blue dot in Fig. 3). As local coordinate frame we use, the Z axis to be vertical to the floor, and X axis to be parallel to the 3D vector from the "right shoulder" to the"left shoulder." For each frame, we use the xyz position relative to the current pelvis joint. Note that in this representation, we ignore the movement of pelvis in the sequence. We also explicitly encode the velocity of joints. Let the i-th joint's position of frame t be J_i^t. Then, the velocity of a joint S_i^t is approximated as the temporal difference between two consecutive frames:

$$S_i^t = (J_i^{t+1} - J_i^t)/\delta t,$$

while δt represents the time interval between consecutive

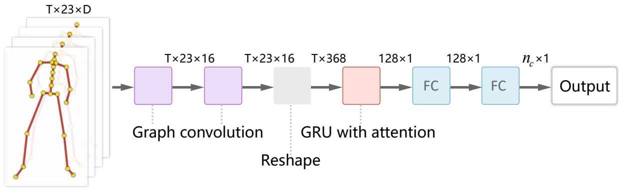


Fig. 4 We represent a skeleton sequence by a 3D tensor of size $T \times J \times D$, T representing the frame length, J the number of joints, and D the feature dimension of each joint, respectively. Our classifier for object property values is made of graph convolution layers, GRU, and fully connected layers. The size of the tensor after each layer is indicated in the figure with n_C denoting the number of classes for an object property, *e.g.*, n_C=6 when the input is a lifting motion and the object property is the weight of box being lifted.

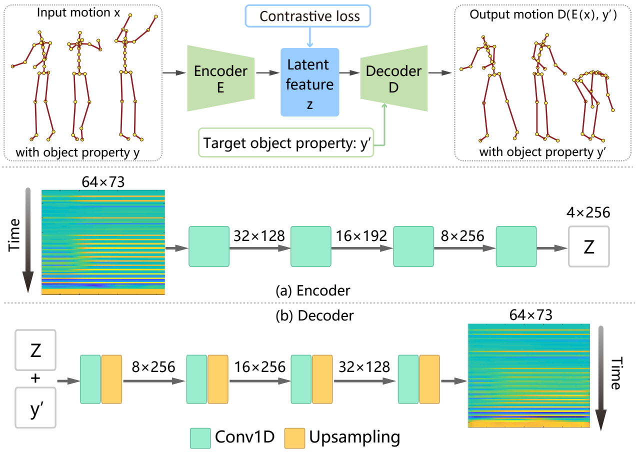


Fig. 5 Network for motion transfer driven by object properties. That is, by changing the object property value y, we may generate human motions that match well with the given property value.

frames.

4.2 Object property classifier

In practice, our object property classifier consists of two graph convolution layers, a GRU layer [7], and then two fully connected (FC) layers for the final classification, i.e., the object property inference; see Fig. 4. The graph convolution layer computes the per-joint features considering the known human body skeleton topology. The GRU layer with attention accumulates the information of all frames and computes the importance of each joint. The combination of graph convolution layers and GRU units enables us to better infer object property values from the same types of motions.

Graph convolution layer. Graph convolution usually deals with the undirected graph. As the skeleton is a hierarchical tree structure, for a given joint, we only consider its parent, instead of all neighbors, to apply a convolution. Formally, for the i-th joint of frame t, its feature after graph convolution $\mathbf{x}'_{t,i}$ is:

$$\mathbf{x}'_{t,i} = \mathbf{Relu}\left(\mathbf{W}_g \begin{bmatrix} \mathbf{x}_{t,i} \\ \mathbf{x}_{t,j} - \mathbf{x}_{t,i} \end{bmatrix} + \mathbf{b}_g\right), \qquad (1)$$

where $\mathbf{x}_{t,i}$ represents the feature of this joint fed to this layer, j is its parent's index, and $\mathbf{W}_g, \mathbf{b}_g$ are the learnable weights for a graph convolution layer. Experiments clearly show that using skeleton topology information can improve the inference accuracy; see e.g., Fig. 9. We use this asymmetric edge function as suggested in [52].

The GRU layer with attention. Attention mechanics is widely used in skeleton-based action recognition. It can improve action recognition and discover the relative importance of joints and frames. For example, Zhang *et al.* [60] use an element-wise attention gate to a RNN block to improve action recognition. We also add a joint-wise gate to the RNN cell. The attention value of each joint of frame t is computed based on the hidden state of the RNN cell $\mathbf{H}_{t-1}$:

$$a_{t,i} = \mathbf{sigmoid}(\mathbf{W}_h\mathbf{H}_{t-1} + \mathbf{W}_x\mathbf{x}_{t,i} + \mathbf{b}_a), \qquad (2)$$

where $\mathbf{x}_{t,i}$ represents the feature of the i-th joint fed to the RNN cell, and $\mathbf{W}_h, \mathbf{W}_x, \mathbf{b}_a$ are the learnable weights for an attention convolution layer. Then, the input fed to the RNN cell is updated as $\tilde{\mathbf{x}}_{t,i} = (1 + a_{t,i})\mathbf{x}_{t,i}$, where $a_{t,i}$ represents the importance of i-th joint at frame t.

Implementation details. For all experiments presented here, we use $J = 23$ major body joints. We use the classic cross entropy loss as it is a classification problem. For skeletal representation, we apply a normalization pre-processing step. The lengths of collected motion samples vary from 3s to 6s. Additionally, we used data augmentation to increase the number of samples and to remove the rotation bias. We rotated each sequence along the Z axis 10 times and cropped 10 sub-sequences from each original and rotated sequence. The rotation angles were drawn from a uniform distribution between $[0, \pi)$, and the cropping ratios were drawn from a uniform distribution $U[0.9, 1]$. This data augmentation enlarged the size of our skeletal motion dataset by 100 times. We down-sample each sub-sequence to 30 frames. We used TensorFlow with the network initialized with Adam optimizer with a batch size of 32 and a learning rate of 0.0001. Training was stopped after 60 epochs by default.

5 Object Property-aware Motion Transfer

In the synthesis content, our goal is to use target object property values to guide motion transfer for a given actor. Given an interaction skeletal motion x whose object property value is y, and a new target object property value

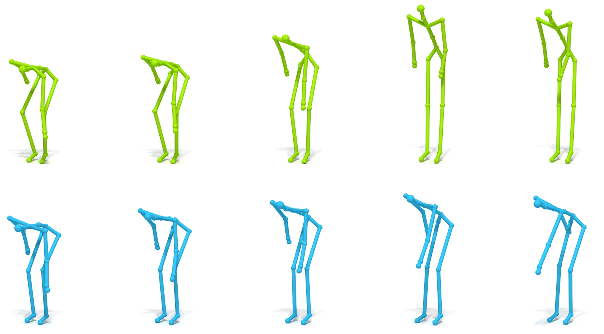


Fig. 6 Given a fishing motion with a long rod (the green), we transfer the rod from *long* to *short* to get a new motion (the blue).

y', we want to generate new skeletal motion x' that matches the given target property value y'.

Inspired by [1, 17], we use an encoder-decoder structure to perform this motion retargeting; see Fig. 5. The encoder E converts an input motion to a latent space $z = E(x)$, and the decoder D synthesizes a new motion conditioned on the target property value, denoted as $D(E(x), y')$. To train the network, we use a loss function consisting of two terms: a reconstruction loss and a contrastive loss.

The *reconstruction loss* aims to constrain the encoder and decoder. We want the output motion to be similar to the motion performed by the same subject under the target property value y', denoted by $\hat{x}$. When y' equals y, $\hat{x}$ equals x. We use the Euclidean loss in the local coordinate frame to measure the quality of the reconstruction:

$$\mathcal{L}_{rec}(E, D) = \mathbb{E}_{x,y'} \|D(E(x), y') - \hat{x}\|_2^2. \qquad (3)$$

The exact choice of the reconstruction loss is not fundamental here. Other reconstruction loss especially designed for motion frames, such as geodesic loss measuring the 3D rotation errors of joints [13], could be used.

Another loss is a *contrastive loss* that ensures that $E(x)$ does not have residual information about the input object property [15]:

$$\begin{aligned}\mathcal{L}_{ctr}(E) =& \mathbb{E}_{x,x^+} \|E(x) - E(x^+)\|_2^2 + \\ & \mathbb{E}_{x,x^-} \left[\alpha - \|E(x) - E(x^-)\|_2\right]_+^2.\end{aligned} \qquad (4)$$

To help disentanglement, we constrain the distance in latent space between different motion samples. Taking an anchor motion x, we compare it with a positive motion x^+ that comes from the same performer under a *different* object property value, and a negative motion x^- that coming from a different performer under the *same* property value. The dissimilarity between the anchor motion and negative motion should be larger than a margin α, and the distance between the anchor motion and positive motion should be small. The full objective functions to optimize the encoder E and decoder D is a combination of two terms:

$$\mathcal{L}(E, D) = \mathcal{L}_{rec}(E, D) + \lambda \mathcal{L}_{ctr}(E), \qquad (5)$$

where λ is a hyper-parameter that controls the relative importance of contrastive loss compared with the reconstruction loss. We use $\lambda = 0.1, \alpha = 5$ in all our experiments.

Here the skeleton sequence for motion transfer is represented by the local and global motion as suggested in [17], which is slightly different from that for object property inference. For local motion, we use joints in XYZ locations of a local frame coordinate, just as the representation for property inference. Global motion consists of the root's global velocity and foot contact labels. See Fig. 5; the rows represent the location of a joint over time. We down-sample the motion to 64 frames.

The encoder is composed of 4 1D convolutional layers with the stride size of two for down-sampling the time axis. The decoder is composed of 4 nearest-neighborhood up-sampling followed by convolution of stride 1 to restore the motion; see Fig. 5.

All models are trained using Adam with $\beta_1 = 0.9$ and $\beta_2 = 0.999$. The batch size is set to 32 for all experiments. We train all models with a learning rate of 0.00001. Training takes about 10 minutes on a server with an Intel Xeon 2.20GHz CPU 10 cores, 256GB memory, and a NVIDIA TitanXP GPU.

6 Results and Evaluation

6.1 Evaluation for object property inference

To measure the model performance on the object property inference, we conducted a cross-subject evaluation. We split the 100 participants into training (60), validation (20),

Tab. 1 Object property inference accuracy (%) on the cross-subject settings. The weight has 6 classes (0, 5, 10, 15, 20, and 25kg). The fragility has 3 levels, implicitly reflected by moving without spill-over an empty bowl, a bowl full of rice and a bowl full of water, respectively. The width of the path, length of the rod, type of the liquid, stiffness of the power twister and water amount in the cup also have 3 levels. The softness of the chair has 4 classes.

Object property	Accuracy (%)	
	Ours	ST-GCN
Lifting a box for weight (6)	**61.8**	57.3
Moving a bowl for fragility (3)	77.5	**78.9**
Walking for path width (3)	**83.9**	73.8
Fishing for length of rod (3)	**80.7**	77.2
Pouring for type of liquid (3)	**62.8**	62.1
Bending for stiffness (3)	**71.6**	44.7
Sitting for softness of chair (4)	**73.7**	66.4
Drinking for water amount inside the cup (3)	**62.5**	57.0

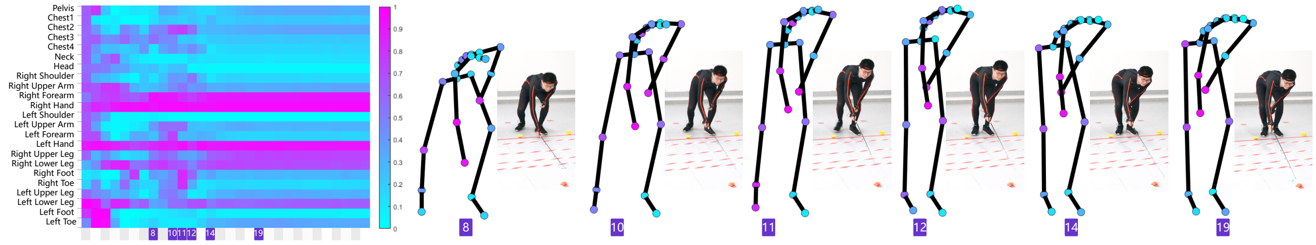


Fig. 7 Estimating the joint-level importance of a fishing motion for inferring the object property. Note that here the color of magenta to cyan indicates the importance from high to low.

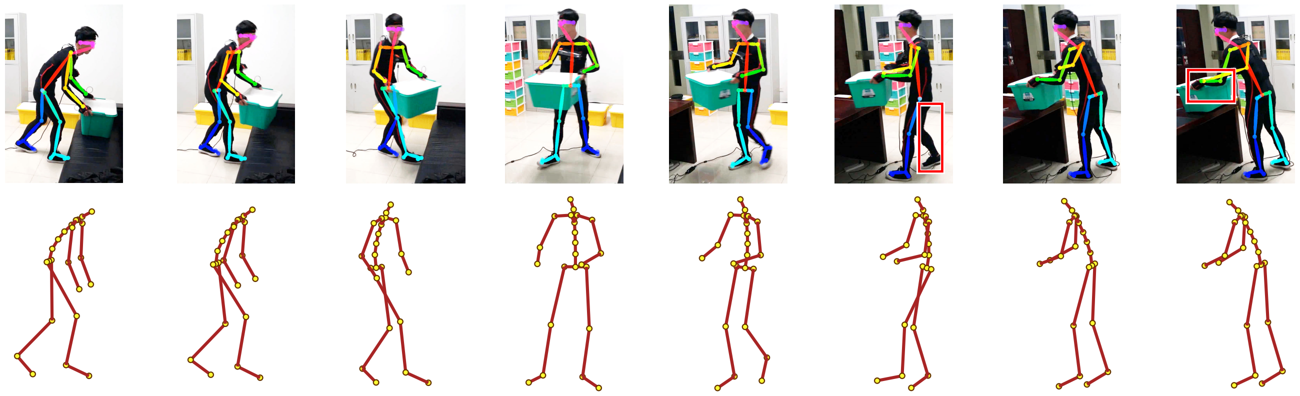


Fig. 8 We show some 2D skeletons extracted from our recorded video at the top, where missing parts are highlighted with red boxes. In comparison, 3D IMU skeletons captured at the corresponding frames are shown underneath, which are clean and complete.

and testing (20) groups, respectively. Hence the testing is done with different people rather than the ones who were employed for training and validation. During training, we select the network parameters with the smallest validation error among all the iterations. Then, we evaluate and report performance on the testing groups.

We implemented several variants to evaluate the impact of different skeleton representations. As using both position and speed achieves the best performance, we applied this representation on other tests. We reported the object property inference accuracy on all eight types of motions. To evaluate, we used a state-of-the-art method for action recognition based on skeletons to set a baseline. We also evaluate the utility of the graph convolution layer and GRU units with attention. Furthermore, we test the inference accuracy regarding the sensitivity of the object property difference.

Tab. 1 shows the object property inference accuracy (%) on the cross-subject settings. The performance looks not very impressive by a first glance at the numbers. Nonetheless, in consideration of the subtle difference among motions under different object properties, we believe this accuracy is reasonable. Furthermore, in most cases, our method outperforms the baseline. We describe the detail of lifting motion in the following as an example. Lifting motion is for the weight estimation from human interaction motions. We trained a classifier that outputs 6 classes corresponding to the weights from 0kg to 25kg with a step of 5kg. The accuracy is about 62% on the cross-subject setting. Considering that the weight difference among the classes are relatively small and the lifting motion is also highly related to the strength of the performer, the resulting estimation accuracy is effective for such subtle changes.

Baseline. We used a state-of-the-art method for action recognition based on skeletons [57] (denoted by ST-GCN) to be a baseline to evaluate the fine-grained motion inference. ST-GCN consists of 9 layers and has about 0.3 million parameters, which is about ten times larger than our model. The original network performed very poorly probably due to the small size of our motion dataset. Setting the layer number as three achieved the best performance during our tuning. We thus reduced the original ST-GCN to three layers. This also leads to a similar parameter setting as ours. We also used both position and speed to represent the skeletal motion. The last column in Tab. 1 shows its performance on the cross-subject setting. Overall speaking, our proposed method has achieved higher

Tab. 2 Impact of different skeleton representations for the inference accuracy (%) on the cross-subject setting.

	Lifting (6)	Walking (3)	Fishing (3)
Position	57.82	76.84	**84.21**
Euler angles	43.38	81.58	73.68
Speed	59.93	79.82	69.4
Angular speed	47.46	73.16	63.51
Position, Euler angles	55.70	79.65	71.58
Position, speed	**61.81**	**83.93**	**80.70**
Position, angular speed	**64.58**	79.47	77.54
Speed, angular speed	55.70	**84.39**	76.49
Speed, Euler angles,	50.56	70.00	66.67
Euler angles, angular speed	56.06	80.53	72.28
Position, Euler angles, angular speed	50.35	78.42	70.18
Position, speed, angular speed	**62.32**	**82.98**	**78.95**
Position, Euler angles, speed	56.55	81.58	71.93
Position, Euler angles, speed, angular speed	58.73	82.98	78.95

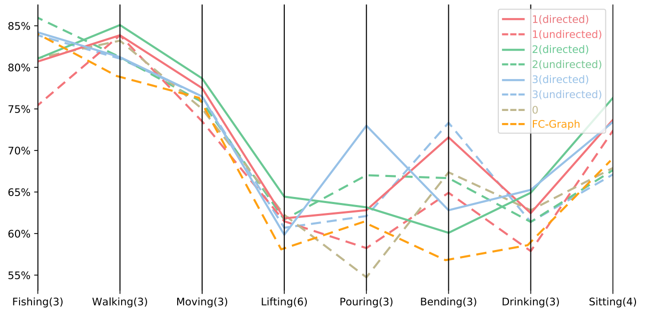


Fig. 9 Parallel coordinates representation for inference accuracy with different ways of computing per-joint features in the 2 graph convolution layers. Each vertical axis represents the inference accuracy from a type of motion. Each line represents a setting. Considering all motions types, it seem good to use the parent of a joint to compute joint feature (the red solid line).

inference accuracy.

Choices of skeleton representation. To evaluate the impact of skeleton representations, we tried several variants. A skeleton sequence was represented by the positions of joints, or the rotation matrix of bones. Similarly, the motion dynamic was measured by the joint speeds or bone angular speeds. We represented the skeleton sequence by different forms, and then evaluated their performance on object property inference of three different motions (i.e., lifting, walking, fishing). All other settings were exactly the same. Tab. 2 shows that the best representation varies for different object properties. Yet overall speaking, using both position and speed is a good option. So this representation was used in other experiments.

Graph convolution. To evaluate the impact of the graph convolution layer regarding per joint feature, we fixed other layers and only changed the two graph convolution layers, and report its performance on object property inference; see Fig. 9. We evaluated on different settings: ignoring the connections between joints and only considering the joint itself to compute per joint feature (similar to PointNet [6]), or treating the skeleton as a tree whose root is the pelvis (directed graph), or treating it as an undirected graph. We also considered different numbers of ancestors (from 1 to 3) of each joint. For an undirected graph, we also considered its k-degree neighborhoods using $k = 1, 2, 3$, or all nodes (FC-Graph) in our tests. Fig. 9 shows that though the inference performance varies across the types of motions, considering a joint's parent to compute its feature is a good option.

Tab. 3 The object weight and water amount inference accuracy (%) under different configurations: two, three, or six classes. See text for more details.

Lifting (kg)				Drinking	
5/25 (2)	10/15 (2)	5/15/25 (3)	(6)	Empty/Full (2)	(3)
94.7	78.7	81.7	61.8	86.8	62.5

Joint-level attention. The learned attentions marginally improved the object property inference, especially for the rod length inference from the Fishing and the softness of chair inference from Sitting motion, both increased about 4%. We visualized the attention weights on joints by the color. For better visualization, we linearly mapped the squared attention values to colors to highlight the importance. Fig. 7 shows the attention weights on the two arms are large for the *fishing* motion, consistent with our human intuition.

Weight and water amount sensitivity. To evaluate the inference accuracy regarding to the sensitivity of the object property difference, we trained and tested the model with several different subsets of motion samples, i.e., using samples with only some specific property values. For example, when evaluating the model's ability to distinguish 5 kg from 10 kg, only motion samples with these two weights were used. All other settings were exactly the same.

Tab. 3 shows that the inference performance is related with the weight label distribution. Note that 2-class classification accuracy drops dramatically from 94.7 down to 78.7 when classifying 10/15kg boxes instead of 5/25kg, even lower than the 3-class classification accuracy of classifying 5/15/25kg. We argue that this is mainly caused by the small dynamic motion difference when lifting boxes are close in weight. The water amount label distribution also shows a similar trend.

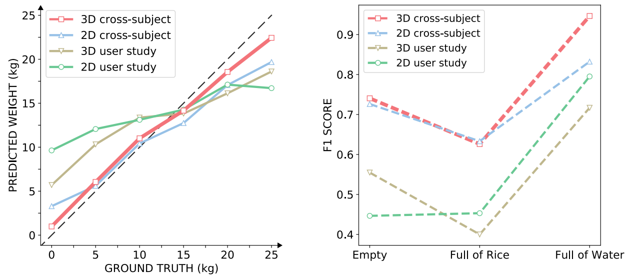


Fig. 10 Left: The average of predicted weights by our model and human observers on both 2D and 3D skeletal cases, where the weights vary among 0, 5, 10, 15, 20, and 25kg. The 6-class weight predictions by our model using 3D skeletons are much closer to ground truth indicated by the slant black line. Right: The F1 score per class of the fragility estimation by our model and human observers. On both 2D and 3D skeletal cases, our method (see the red and blue marks) achieves better results.

Tab. 4 When adding rendered skeletons into training, the object inference accuracy (%) (such as the weight by lifting and the fragility by moving) from videos can be improved significantly as compared below.

	Lifting (6)	Moving (3)
without	51.6	62.9
with	**61.4**	**71.4**

6.2 Comparison with videos

Property inference from only videos. We additionally evaluate the weight and fragility inference performance from different input sources. In particular, we have tested the performance using 2D skeleton sequences directly extracted from videos that were recorded from a fixed view. We used OpenPose detector [5] to extract 25 body keypoints in 2D to get image-space skeletons using videos. Due to the fixed camera view and the occlusion of interacting objects, extracted 2D skeletons may have large missing parts in some frames; see e.g., Fig. 8 (top). We choose the most representative 17 body joints, and replace the 3D IMU skeletons with corresponding 2D video skeletons. Now the skeleton sequences have only x and y positions without z dimension. The speed and acceleration attributes are not used as there are unavoidable flickers in video sequences and they cannot be easily lifted to 3D.

Fig. 10 presents the evaluation of 6-class weight classification and 3-class fragility inference on cross-subject settings, by our model trained on 2D and 3D skeletons and human observers. Using 2D skeletons instead of 3D causes some drop in inference accuracy in both weight and fragility estimation, see the red and blue lines. We believe this is mainly due to joint estimation errors, depth information missing, and kinematic flicker artifacts.

Property inference from videos enhanced by 3D skeletons. The small size of unoccluded 3D skeletons motion samples may generate thousands of rendered 2D skeletons. Here we show these 2D projections of 3D data can effectively improve the performance of property value estimation from 2D videos. We generated these virtual 2D samples by projecting the 3D joint positions of 3D skeleton sequences according to different camera view angles. For the virtual camera setting, we used a weak-perspective camera model, as suggested by [1], which generates 2D projections of synthetic 3D skeleton sequences. For every 3D sequence, we used 8 fixed views, placed a camera every 22.5 degrees around the actor (covered about 180 degrees in total), and all cameras were set to be horizontal (pitch angle equals to 0).

Tab. 4 presents the evaluation of 6-class weight classification and 3-class fragility inference on the cross-subject setting, by our models trained on 2D skeletons extracted from videos only, or on 2D extracted skeletons and rendered 3D skeletons. The trained models were tested only on 2D extracted skeletons. In the second case, The ratio of extracted and rendered skeletons was 1 : 8. Clearly using additional virtual skeletons can effectively improve the performance.

6.3 Evaluation for property-aware motion transfer

We again split the 100 subjects into training (60), validation (20), and test (20) groups, respectively. During training, we select the network parameters with the smallest validation error among all the iterations. We evaluate and report performance on the test groups.

Latent space visualization. Fig. 13 shows the latent space of motion samples after projecting the latent features to a 2D image using t-SNE. Each point represents a motion sample of a subject lifting a 0 kg or 25 kg box. The leftmost figure shows that they are clustered according to object property values without contrastive loss. This is due to the motion differences among different subjects are smaller than that of lifting 0 kg and 25 kg boxes. With the contrastive loss, the features start to disentangle from object properties and become more related to the subjects.

Results. Fig. 11, 12, and 14 show three generated motions by changing the object property values. Please also refer to the supplementary video for more examples. When the input is a walking motion on a width path by an unseen subject, we transfer motion to walk on a narrow path, like a catwalk

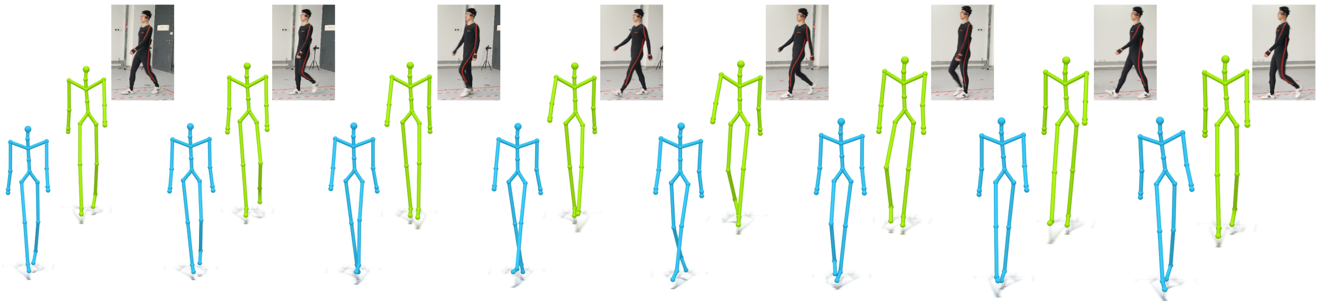


Fig. 11 Given a motion sequence of an unseen subject walking on the wide path (in green), we can generate a new sequence that looks like the subject was walking on a narrow path (in blue).

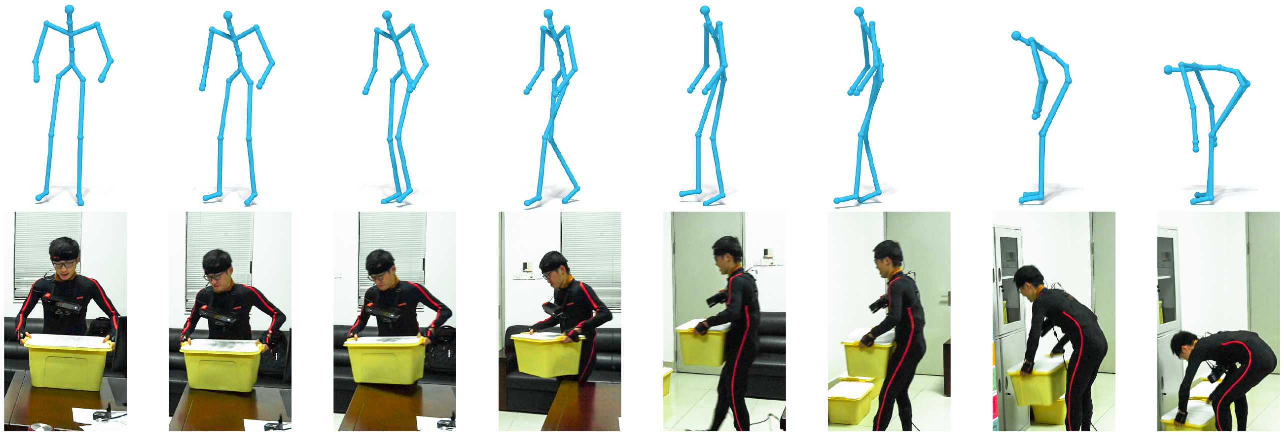
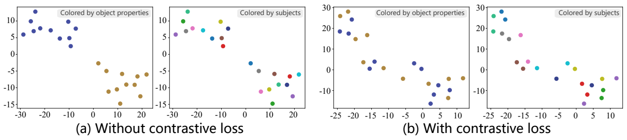


Fig. 12 Given the motion sequence shown in Fig. **??**, we can generate a new sequence that looks like the subject was lifting a heavy box, but it was too heavy to be lifted. The generated motion is similar to the ground truth as shown with a sequence of RGB images at the bottom.



(a) Without contrastive loss

(b) With contrastive loss

Fig. 13 Latent variables after encoder of several lifting motions with 0 kg and 25 kg boxes are projected to 2D space. Without contrastive loss (a), the left is colored by object properties, and the right by subjects. With contrastive loss (b), colored the same way.

model. Given a motion sequence of an unseen subject lifting a light box from a table to a closet, we generate a new sequence that looks like the box is too heavy to be lifted up; see Fig. 12.

In Fig. 14, we show a generated sequence that drinking from an empty cup, given an unseen motion sequence drinking from a cup full of water using two hands. As the unseen motion is considerably different from the training set, the generated motion deviates from the input. However, sometimes it is ambiguous what is the correct motion. Note during training, we constrain the synthesized motion conditioned on a target property value to be similar to the motion performed by the same subject of given object property. Multiple options may likely match the desired motion property value. It would be desirable if we could synthesize the one that is most similar to the input motion.

6.4 User study

We conducted two user studies. The first one is to investigate a human observer's perception on the weight and fragility inference from skeleton sequences. We considered both the 3D skeletons captured and 2D skeletons extracted from videos. The second user study is conducted to evaluate the property-aware motion transfer on the sitting and walking sequences.

The first user study. In the study, a test consisted of watching a video of skeletal motion of an actor lifting a box or moving a bowl, then predicting the unseen object's property by choosing an answer from multiple choices. For LiftingBox sequence, six choices were provided: 0, 5, 10, 15, 20, and 25kg. For MovingBowl sequence, three

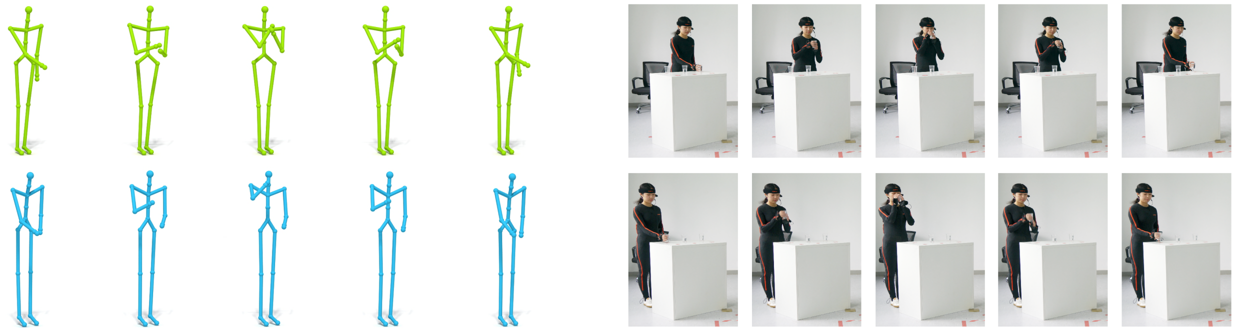


Fig. 14 Given an unseen motion sequence of drinking from a cup full of water using two hands, we generate a new sequence that drinking from an almost empty cup (the blue skeletons in second row). In the training set, all the subjects drink water using one hand. The corresponding RGB images of the actor are shown in the right for a better illustration.

choices were provided: empty, fully filled with rice, and fully filled with water. There were a total of 12 tests. To help answering the questions, 4 demos with correct answers were played before the tests started. These motion samples were randomly chosen from the testing group. Each video was about 3–6 seconds long. All participants had full control over these videos, e.g., start, pause, stop and navigate in time, etc. A total of 60 participants were recruited. Each participant did the user study twice. The first time they predicted the weight from videos of rendered 3D skeletons, and the second time they predicted the weight from 2D video skeletons. Note that 2D video skeletons have large missing parts in some frames due to the occlusions introduced by human body shape or the objects, while the rendered ones have much fewer occlusion cases caused by bones. These skeletons were drawn with the same color encoding. The total study time for each participant was around 10 minutes.

Fig. 10 (left) shows the average predicted weights by users and our model for boxes of different physical weights. The estimated weights by our model using 3D skeletons as input are much closer to the physical ground truth than other settings. Note that our reported human performance is slightly lower than that reported in Runeson and Frykholm's work [40]. A possible reason is that a smaller weight step (5kg) and more weight classes (6) were used in our user study. Fig. 10 (right) displays the F1 scores of user study and our model on the fragility inference. Note it is challenging to distinguish an empty bowl from a bowl full of rice, but still, our model outperformed on both 2D and 3D skeletal cases.

The second user study. A total of 60 participants were recruited and divided into two groups, watching the sitting and walking sequences, respectively. Every participant did 12 tests, and 4 demos with correct answers were played before the tests started. A test contained two parts. The first task is to judge if the given motion was synthesized or captured. The second task is to select the associated object property of the given motion, while only 2 choices were provided. For example, to select the path being walked on was wide or narrow, or the chair being sit on was soft or hard. Other settings are similar to the first user study. Fig. 15 shows the performance of participants on motion source and object property inference. The lightness of a square encodes the number of participants with a particular inference accuracy, the darker the higher. For majority participants, the source inference accuracy is about 60%, while the property inference accuracy is above 90%, indicating that our synthesized motions are quite close to real captured ones.

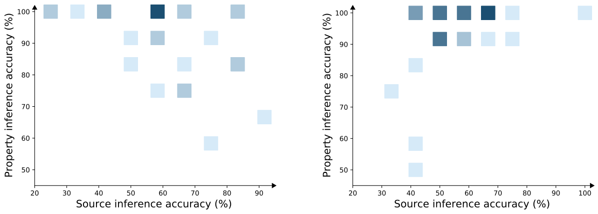


Fig. 15 The scatter map of participants' accuracy (%) on guess the motion's source (synthesized or captured), and on the object property inference. Left: sitting; right: walking. Shades of blue indicate the number of participants, darker being higher.

7 Conclusions and Future Work

The primary goal of this work is to study human interaction motions represented by skeleton sequences, and investigate whether and how well a machine can learn to infer the properties of unseen interacting objects, and to what extent we can have control on the synthesis of motions with target object properties. We have built up a large multi-modal dataset for such object property inference from fine-grained human interaction motions with 4,000+ samples, which consist of 100 participants performing 8 different tasks, and thus related to 8 different object properties.

Using 3D skeleton sequences alone, we have learned to infer the properties of interacting objects by treating it as a classification problem, and evaluated our trained model in various settings. The collected 3D skeleton sequences allows data-driven learning, and help achieve better inference accuracy in comparison with using other data sources or even human observers. We have presented a network to disentangle object property from the motion. The disentangling, in turn, allows the synthesis of modified motion with a target object property. This control over the actions enriches the dataset on one hand, and optimizes the specific animation of particular individuals on the other.

Limitations. Due to the design, our target problem is limited in the defined scenarios with pre-defined human motions and object properties. The inference and transfer tasks are solved separately, while exploiting features extracted during inference to guide the synthesis part might be possibly better. The main techniques used in both the inference and transfer tasks are well established.

Separate classifiers have to be trained for different type of motions, and the accuracy is not that high. We focus only on the intra-class characteristics for the object property inference, but it might be better to address action recognition and object property inference altogether, as the action types provide more global content information.

The object property-aware motion transfer employs an encoder-decoder structure with 1D convolution layers, which might not fully capture the spatial-temporal information of more human motions, in particular, the complex ones. More advanced network structures, such as STRNN [51], could be used to better transfer in-between independent actions.

Future work. Exciting research directions lay ahead as we are only starting to exploit the collected motion data. We have made a large-scale interaction dataset public. We believe that this dataset will stimulate further research, and in the future, we will strive not only to increase the number of samples, but also the types of human-object interactions. Previous works have shown that some other properties, e.g., size and geometric shape, are quite hard to be estimated from a pantomimed action [50]. To be able to deal with more diverse object properties, we are also considering fusing more visual inputs, e.g., videos and depth sequences, with 3D skeletal motions.

Another promising direction is to discover exactly which parts of the skeleton are critical for the specific object property inference, by considering more sophisticated attention models or computing more advanced skeletal features. Further exploration could also focus on designing new networks that can learn and encode skeletal motions in a learned latent space, instead of being explicitly provided parameterization. It is certainly more exciting if we can directly predict object properties from 2D video inputs of large occlusions with high accuracy using a trained model on 3D skeletal motions, eventually leading to new modes of authoring video sequences.

Acknowledgements

We sincerely thank the reviewers for their valuable comments. This work was supported in parts by Shenzhen Innovation Program (JCYJ20180305125709986), NSFC (61861130365, 61761146002), GD Science and Technology Program (2020A0505100064, 2015A030312015), and DEGP Key Project (2018KZDXM058).

References

[1] K. Aberman, R. Wu, D. Lischinski, B. Chen, and D. Cohen-Or. Learning character-agnostic motion for motion retargeting in 2d. *ACM Trans. on Graphics (Proc. of SIGGRAPH)*, 38(4):75, 2019.

[2] M. Andriluka, U. Iqbal, E. Insafutdinov, L. Pishchulin, A. Milan, J. Gall, and B. Schiele. Posetrack: A benchmark for human pose estimation and tracking. In *CVPR*, 2018.

[3] A. Aristidou, D. Cohen-Or, J. K. Hodgins, Y. Chrysanthou, and A. Shamir. Deep motifs and motion signatures. *ACM Trans. on Graphics (Proc. of SIGGRAPH Asia)*, 38(6):187:1–187:13, 2018.

[4] R. Blake and M. Shiffrar. Perception of human motion. *Annual Review of Psychology*, 58(1):47–73, 2007.

[5] Z. Cao, T. Simon, S.-E. Wei, and Y. Sheikh. Realtime multi-person 2d pose estimation using part affinity fields. In *CVPR*, 2017.

[6] R. Q. Charles, H. Su, M. Kaichun, and L. J. Guibas. PointNet: Deep learning on point sets for 3d classification and segmentation. In *CVPR*, jul 2017.

[7] K. Cho, B. Van Merriënboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio. Learning phrase representations using rnn encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078*, 2014.

[8] CMU. Carnegie mellon university mocap database, 2018.

[9] J. W. Davis and H. Gao. Recognizing human action efforts: an adaptive three-mode PCA framework. In *ICCV*, 2003.

[10] A. F. de C. Hamilton, D. W. Joyce, J. R. Flanagan, C. D. Frith, and D. M. Wolpert. Kinematic cues in perceptual weight judgement and their origins in box lifting. *Psychological Research*, 71(1):13–21, 2005.

[11] G. Gkioxari, R. Girshick, P. Dollár, and K. He. Detecting and recognizing human-object intaractions. *CVPR*, 2018.

[12] H. Grabner, J. Gall, and L. V. Gool. What makes a chair a chair? In *CVPR*, 2011.

[13] L.-Y. Gui, Y.-X. Wang, X. Liang, and J. M. F. Moura. Adversarial geometry-aware human motion prediction. In *ECCV*, pages 823–842. 2018.

[14] A. Gupta and L. S. Davis. Objects in action: An approach for combining action understanding and object perception. In *CVPR*, 2007.

[15] R. Hadsell, S. Chopra, and Y. LeCun. Dimensionality reduction by learning an invariant mapping. In *CVPR*, volume 2, pages 1735–1742, June 2006.

[16] E. S. L. Ho, T. Komura, and C.-L. Tai. Spatial relationship preserving character motion adaptation. *ACM Trans. on Graphics (Proc. of SIGGRAPH)*, 29(4), July 2010.

[17] D. Holden, J. Saito, and T. Komura. A deep learning framework for character motion synthesis and editing. *ACM Trans. on Graphics (Proc. of SIGGRAPH)*, 35(4):1–11, 2016.

[18] E. Hsu, K. Pulli, and J. Popović. Style translation for human motion. *ACM Trans. on Graphics*, 24(3):1082–1089, July 2005.

[19] R. Hu, M. Savva, and O. van Kaick. Functionality representations and applications for shape analysis. *Computer Graphics Forum (Proc. of Eurographics)*, 37(2):603–624, 2018.

[20] R. Hu, Z. Yan, J. Zhang, O. V. Kaick, A. Shamir, H. Zhang, and H. Huang. Predictive and generative neural networks for object functionality. *ACM Trans. on Graphics (Proc. of SIGGRAPH)*, 37(4):1–13, 2018.

[21] E. Insafutdinov, L. Pishchulin, B. Andres, M. Andriluka, and B. Schiele. DeeperCut: A deeper, stronger, and faster multi-person pose estimation model. In *ECCV*, October 2016.

[22] Y. Jiang, H. Koppula, and A. Saxena. Hallucinated humans as the hidden context for labeling 3d scenes. In *CVPR*, pages 2993–3000, 2013.

[23] Y. Jiang, H. Koppula, and A. Saxena. Modeling 3d environments through hidden human context. *IEEE Trans. Pattern Anal. Mach. Intell.*, 38(10):2040–2053, 2016.

[24] A. Kanazawa, M. J. Black, D. W. Jacobs, and J. Malik. End-to-end recovery of human shape and pose. In *CVPR*, 2018.

[25] C. Kang and S.-H. Lee. Scene reconstruction and analysis from motion. *Graphical Models*, 94:25–37, 2017.

[26] K. Kato, Y. Li, and A. Gupta. Compositional learning for human object interaction. In *ECCV*, October 2018.

[27] Q. Ke, M. Bennamoun, S. An, F. Sohel, and F. Boussaid. A new representation of skeleton sequences for 3d action recognition. In *CVPR*, 2017.

[28] V. G. Kim, S. Chaudhuri, L. Guibas, and T. Funkhouser. Shape2pose: human-centric shape analysis. *ACM Trans. on Graphics (Proc. of SIGGRAPH)*, 33(4):1–12, 2014.

[29] C. Li, Q. Zhong, D. Xie, and S. Pu. Co-occurrence feature learning from skeleton data for action recognition and detection with hierarchical aggregation. In *Proc. Int. Joint Conf. on Artificial Intelligence*, page 786–792, 2018.

[30] X. Li, S. Liu, K. Kim, X. Wang, M.-H. Yang, and J. Kautz. Putting humans in a scene: Learning affordance in 3d indoor environments. In *CVPR*, 2019.

[31] C. Liu, Y. Hu, Y. Li, S. Song, and J. Liu. Pku-mmd: A large scale benchmark for continuous multi-modal human action understanding. *arXiv preprint arXiv:1703.07475*, 2017.

[32] J. Liu, A. Shahroudy, D. Xu, and G. Wang. Spatio-temporal LSTM with trust gates for 3d human action recognition. In *ECCV*, October 2016.

[33] J. Liu, G. Wang, P. Hu, L.-Y. Duan, and A. C. Kot. Global context-aware attention LSTM networks for 3d action recognition. In *CVPR*, 2017.

[34] L. Lo Presti and M. La Cascia. 3d skeleton-based human action classification. *Pattern Recogn.*, 53(C):130–147, May 2016.

[35] D. Mehta, S. Sridhar, O. Sotnychenko, H. Rhodin, M. Shafiei, H.-P. Seidel, W. Xu, D. Casas, and C. Theobalt. VNect: real-time 3d human pose estimation with a single rgb camera. *ACM Trans. on Graphics (Proc. of SIGGRAPH)*, 36(4):1–14, 2017.

[36] A. Monszpart, P. Guerrero, D. Ceylan, E. Yumer, and N. J. Mitra. imapper: Interaction-guided joint scene and human motion mapping from monocular videos. *ACM Trans. on Graphics (Proc. of SIGGRAPH)*, 38(4), July 2019.

[37] A. Newell, K. Yang, and J. Deng. Stacked hourglass networks for human pose estimation. In *ECCV*, October 2016.

[38] G. Pavlakos, X. Zhou, and K. Daniilidis. Ordinal depth supervision for 3d human pose estimation. In *CVPR*, 2018.

[39] J. Podda, C. Ansuini, R. Vastano, A. Cavallo, and C. Becchio. The heaviness of invisible objects: Predictive weight judgments from observed real and pantomimed grasps. *Cognition*, 168:140–145, 2017.

[40] S. Runeson and G. Frykholm. Visual perception of lifted weight. *Journal of Experimental Psychology: Human Perception and Performance*, 7(4):733, 1981.

[41] A. G. Rıza, N. Neverova, and I. Kokkinos. Densepose: Dense human pose estimation in the wild. In *CVPR*, 2018.

[42] M. Savva, A. X. Chang, P. Hanrahan, M. Fisher, and M. Nießner. SceneGrok: inferring action maps in 3d environments. *ACM Trans. on Graphics (Proc. of SIGGRAPH)*, 33(6):1–10, 2014.

[43] F. Schmidt, V. C. Paulun, J. J. R. van Assen, and R. W. Fleming. Inferring the stiffness of unfamiliar objects from optical, shape, and motion cues. *Journal of Vision*, 17(3):18–18, March 2017.

[44] A. Shahroudy, J. Liu, T.-T. Ng, and G. Wang. NTU RGB+D: A large scale dataset for 3d human activity analysis. In *CVPR*, 2016.

[45] Y. Shen, L. Yang, E. S. L. Ho, and H. P. H. Shum. Interaction-based human activity comparison. *IEEE Trans. Visualization & Computer Graphics*, pages 1–1, 2019.

[46] S. Song, C. Lan, J. Xing, W. Zeng, and J. Liu. An end-to-end spatio-temporal attention model for human action recognition from skeleton data. In *AAAI Conf. on Artificial Intelligence*, 2017.

[47] T. A. Stoffregen and S. B. Flynn. Visual perception of support-surface deformability from human body kinematics. *Ecological Psychology*, 6(1):33–64, 1994.

[48] B. Tekin, A. Rozantsev, V. Lepetit, and P. Fua. Direct prediction of 3d body poses from motion compensated sequences. In *CVPR*, 2016.

[49] D. Tome, C. Russell, and L. Agapito. Lifting from the deep: Convolutional 3d pose estimation from a single image. In *CVPR*, 2017.

[50] L. M. Vaina, H. Goodglass, and L. Daltroy. Inference of object use from pantomimed actions by aphasics and patients with right hemisphere lesions. *Synthese*, 104(1):43–57, 1995.

[51] H. Wang, E. S. L. Ho, H. P. H. Shum, and Z. Zhu. Spatio-temporal manifold learning for human motions via long-horizon modeling. *IEEE Trans. Visualization & Computer Graphics*, pages 1–1, 2019.

[52] Y. Wang, Y. Sun, Z. Liu, S. E. Sarma, M. M. Bronstein, and J. M. Solomon. Dynamic graph cnn for learning on point clouds. *ACM Trans. on Graphics*, 2019.

[53] S.-E. Wei, V. Ramakrishna, T. Kanade, and Y. Sheikh. Convolutional pose machines. In *CVPR*, 2016.

[54] J. Wu, J. Lim, H. Zhang, J. Tenenbaum, and W. Freeman. Physics 101: Learning physical object properties from unlabeled videos. pages 39.1–39.12, 2016.

[55] J. Wu, I. Yildirim, J. J. Lim, W. T. Freeman, and J. B. Tenenbaum. Galileo: Perceiving physical object properties by integrating a physics engine with deep learning. page 127–135, 2015.

[56] S. Xia, C. Wang, J. Chai, and J. Hodgins. Realtime style transfer for unlabeled heterogeneous human motion. *ACM Trans. on Graphics (Proc. of SIGGRAPH)*, 34(4), July 2015.

[57] S. Yan, Y. Xiong, and D. Lin. Spatial temporal graph convolutional networks for skeleton-based action recognition. In *AAAI Conf. on Artificial Intelligence*, 2018.

[58] B. Yao and L. Fei-Fei. Modeling mutual context of object and human pose in human-object interaction activities. In *CVPR*, 2010.

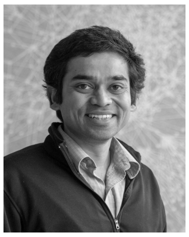
[59] M. E. Yumer and N. J. Mitra. Spectral style transfer for human motion between independent actions. *ACM Trans. on Graphics (Proc. of SIGGRAPH)*, 35(4), July 2016.

[60] P. Zhang, J. Xue, C. Lan, W. Zeng, Z. Gao, and N. Zheng. Adding attentiveness to the neurons in recurrent neural networks. In *ECCV*, October 2018.

Qian Zheng received the doctoral degree in computer science from Shenzhen Institutes of Advanced Technology (SIAT), Chinese Academy of Sciences, in 2015. She is an assistant professor in college of computer science and software engineering, Shenzhen University. Her interests include computer graphics and information visualization.

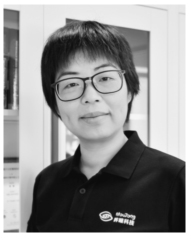
Weikai Wu is a software engineer in TCL. He received his MS in Computer Science from Shenzhen University in 2020.

Hanting Pan is a software engineer in Orbbec. He received his MS in Computer Science from Shenzhen University in 2020.



Niloy Mitra leads the Smart Geometry Processing group in the Department of Computer Science at University College London. He received his PhD from Stanford University under the guidance of Leonidas Guibas. His research interests include shape analysis, creativeAI, and computational design and fabrication. Niloy received the Eurographics Outstanding Technical Contributions Award in 2019, the BCS Roger Needham award in 2015, and the ACM Siggraph Significant New Researcher Award in 2013.

Daniel Cohen-Or is a Professor in the School of Computer Science. He received his Ph.D. from the State University of New York at Stony Brook in 1991. He was the recipient of Eurographics Outstanding Technical Contributions Award in 2005, and ACM SIGGRAPH Computer Graphics Achievement Award in 2018. In 2019 he won Kadar Family Award for Outstanding Research. In 2020 he received Eurographics Distinguished Career Award. His research interests are in Computer Graphics, in particular, synthesis, processing and modeling techniques.



Hui Huang is a Distinguished TFA Professor of Shenzhen University, where she directs the Visual Computing Research Center. She received her PhD in Applied Math from The University of British Columbia in 2008. Her research interests span on Computer Graphics, 3D Vision and Visualization. She is currently a Senior Member of IEEE/ACM/CSIG, a Distinguished Member of CCF, and is on the editorial board of ACM Trans. on Graphics and Computers & Graphics.

A Interaction motion dataset collection

WALKING. The experiment on Walking aims for estimating *the width* of the path. Each subject was asked to walk back and forth on three straight paths of different widths. We simulated the width of a path using line markers to indicate path borders, and asked the subjects do not cross the borders. So we have a total of $3 \times 2 \times 100 = 600$ motion samples.

FISHING. The experiment on Fishing aims for estimating *the length* of a fishing rod. Each subject was asked to use a fishing rod to fetch a magnetic object placed in front. The object would attach to the rod's end when being touched. Each subject did 3 trails, with fishing rods of three different lengths. We have a total of $3 \times 3 \times 100 = 900$ motion samples.

POURING. The experiment on Pouring aims for estimating *the type* of liquid. Each subject was asked to pour liquid from a cup to other one. Each subject did 3 trails with three different substances (water, shampoo, and rice). The pouring motions were effected by the viscosity or particle granularity.

BENDING. The experiment on Bending aims for estimating *the stiffness* of a power twister. Each subject was asked to bend a power twister with three different setting, from easy to hard mode.

SITTING. The experiment on Sitting aims for estimating *the softness* of a chair being sat on. Each subject was asked to sit on four chairs of same height but different softness. The hardest chair is made of plastic, and the softest one is a yoga ball.

DRINKING. The experiment on drinking aims for estimating *the amount* of water inside a cup. Each subject was asked to take a cup from a table and get a sip of water. Each subject did 3 trails while the amount of water in the cup changed from almost full, to half full, and to almost empty.

LIFTINGBOX. The experiment on LiftingBox aims to estimate *the weight* of an object from the human motion interaction. Each subject was asked to perform four different tasks in a row: (i) lifting a box from the ground to a sofa; (ii) lifting the box from the sofa to a table; (iii) lifting the box from the table to the top of a closet; finally (iv) putting the box back to the floor. Without letting the subject know, the weight of the carrying box was randomly changed by putting different weight plates into the concealed box, ranging from 0kg to 25kg in a step of 5kg. That is, each subject needed to do 6 trails and did not know if he/she would lift a heavy or light box before each trial, so all the captured motions are naturally close to what happens in our real life. This lifting experiment provides us 1343 motion samples in total, all annotated with the specific task and weight. When a subject failed to lift up a heavy box to somewhere high, he/she did not need to perform the following tasks along the line with the same weight.

MOVINGBOWL. The experiment on MovingBowl aims to judge *the fragility* of an object from human motion interactions. While the weight belongs to a physical property, the fragility leans more to an empirical property. Each subject was asked to perform the similar four tasks in a row as described above, but to move a bowl this time rather than lifting a box. Three same uncovered bowls were used: one empty, one fully filled with rice, and one fully filled with water. That is, each subject was needed to do 3 trails and saw clearly the different states of these three bowls. They were all required to try their best to move the bowls without any spillage. We expect this to capture how cautious the subject was for the target task and how much that correlates to his/her motion in the corresponding trial. The degree of caution should be the highest when moving a bowl full of water, and the lowest when moving an empty bowl, which in turn relates to the level of fragility of an object. All action samples are annotated with one of the three levels of interacting object fragility.