International Journal of Computer Vision (2020) 128:220–239
https://doi.org/10.1007/s11263-019-01234-9

Predicting Intentions from Motion: The Subject-Adversarial Adaptation Approach

Andrea Zunino[1] · Jacopo Cavazza[1] · Riccardo Volpi[1] · Pietro Morerio[1] · Andrea Cavallo[2,4] · Cristina Becchio[2,4] · Vittorio Murino[1,3]

Received: 12 February 2018 / Accepted: 9 September 2019 / Published online: 18 September 2019

Abstract

This paper aims at investigating the action prediction problem from a pure kinematic perspective. Specifically, we address the problem of recognizing future actions, indeed human intentions, underlying a same initial (and apparently unrelated) motor act. This study is inspired by neuroscientific findings asserting that motor acts at the very onset are embedding information about the intention with which are performed, even when different intentions originate from a same class of movements. To demonstrate this claim in computational and empirical terms, we designed an ad hoc experiment and built a new 3D and 2D dataset where, in both training and testing, we analyze a same class of grasping movements underlying different intentions. We investigate how much the intention discriminants generalize across subjects, discovering that each subject tends to affect the prediction by his/her own bias. Inspired by the domain adaptation problem, we propose to interpret each subject as a domain, leading to a novel subject adversarial paradigm. The proposed approach favorably copes with our new problem, boosting the considered baseline features encoding 2D and 3D information and which do not exploit the subject information.

Keywords Action recognition and prediction · Human intentions · Grasping · Kinematic analysis · Adversarial domain adaptation

1 Introduction

Recognizing human actions is an active area of research which is faced under different paradigms in computer vision and pattern recognition. The most straightforward one, namely action recognition, typically consists in the classification of a *fully observed* activity[1] from a video sequence. *Early* activity recognition aims at recognizing an action *before* it is fully disclosed, i.e., from the onset of that same action. Action prediction instead refers to the classification of future actions considering all the events occurring up to a certain time instant (Chakraborty and Roy-Chowdhury 2014).

In this paper, as a different paradigm, we introduce a new and more demanding problem, consisting in the anticipation of future, never observed, actions where the only available input data consists of motion segments of a non-discriminant and apparently unrelated (with respect to the future action) motor act. In other words, the problem we would like to address consists in the prediction of human *intentions*, defined as the overarching goal embedded in an action sequence (see Fig. 1).

Predicting intentions is paramount in many aspects of our social life, as well as for security tasks. For instance, it would be desirable to predict the intention of a person in his/her car stopped at a police checkpoint whether, while opening the

Communicated by Kristen Grauman.

Andrea Zunino and Jacopo Cavazza have contributed equally to this work.

Electronic supplementary material The online version of this article (https://doi.org/10.1007/s11263-019-01234-9) contains supplementary material, which is available to authorized users.

✉ Andrea Zunino
andrea.zunino@iit.it

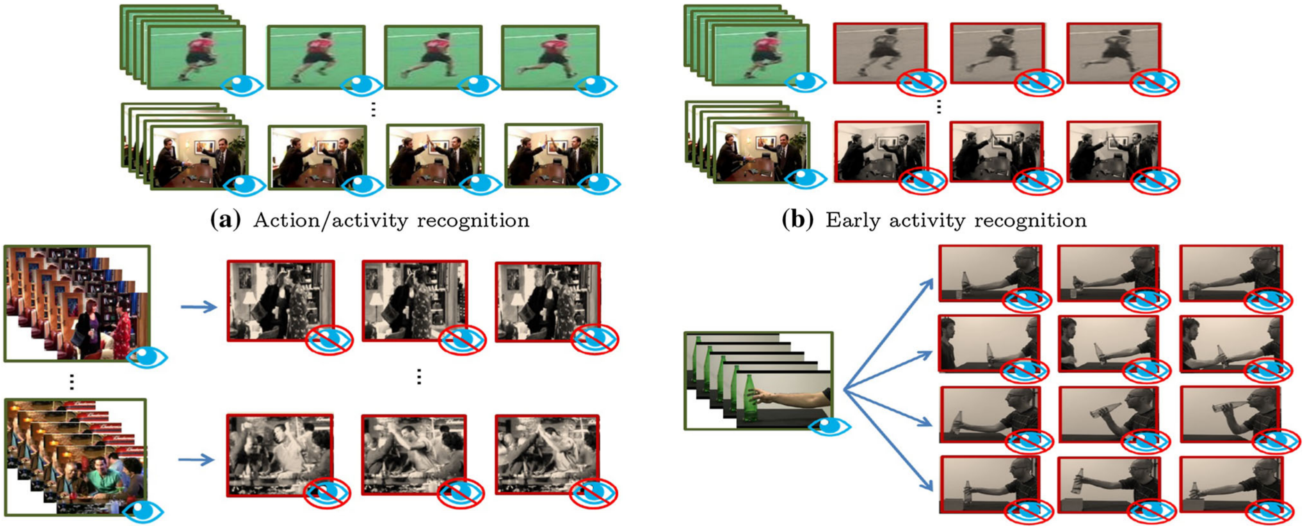
[1] Pattern Analysis and Computer Vision (PAVIS), Istituto Italiano di Tecnologia (IIT), Genoa, Italy

[2] C'MON Cognition, Motion and Neuroscience, Istituto Italiano di Tecnologia (IIT), Genoa, Italy

[3] Department of Computer Science, University of Verona, Verona, Italy

[4] Department of Psychology, University of Torino, Turin, Italy

[1] If not differently specified, activity and action are here used interchangeably.



(a) Action/activity recognition

(b) Early activity recognition

(c) Action prediction

(d) The proposed paradigm of intention prediction

Fig. 1 Four different paradigms. **a** Action/activity recognition: the full sequences is exploited for classification (the top sequence shows the "running" class and the bottom one represents the "high-five" class). **b** Early activity recognition: a few initial frames are observed and classification rely upon such incomplete information (same classes as in **a**). **c** Action prediction: future actions are predicted on the basis of all past events which are class-specific. For instance, in the top sequence a standing up activity leads to predict a "kissing", and, in the bottom, a conversation between a group of friends anticipates a "high-five". **d** Intention prediction: the same class of motor act (at left in the picture, grasping) is analyzed to explain why the motor act itself has been displayed, predicting its underlying intention (at right in the picture and from top to bottom, Pouring, Passing, Drinking and Placing)

glove compartment, it is going to pick documents or taking a gun. Or to detect the intention of a subject standing in front of a bank counter and grabbing something from the pocket, whether it will pick his wallet to deposit money or extract a weapon to attempt a robbery. Further, in social robotics, the capability to predict intentions would enhance the robot interaction performance through a more realistic engagement with humans.

This study is inspired by recent findings from behavioral neuroscience (Ansuini et al. 2014, 2015; Cavallo et al. 2016; Koul et al. 2018, 2019; Becchio et al. 2018; Soriano et al. 2018; Zunino et al. 2018) which assert that the execution of a motor act (e.g., grasping an object) is not solely determined by bio-mechanical constraints imposed by the object that is involved in the interaction and by its intrinsic/extrinsic properties. But, in fact, it also depends on the agent's intention which tends to adapt the current motor act to better fulfill the intention which originated it (e.g., grasp an object to pass to someone or to use it directly Ansuini et al. 2014, 2015).

The difference with respect to the other paradigms in the current computer vision research is subtle but clear: in our case, intentions can be predicted not only using discriminant previous information extracted from a certain anticipative data stream *related* to the action to be performed, but they can also be inferred from the motion of an anticipative *apparently unrelated* action. Apparently unrelated means here that this initial action is not executed for a specific, unique purpose, but it may proceed with different future actions. More specifically, our challenge lies in predicting future different intentions originating from the *same class* of anticipative motor acts (e.g., grasping an object for different usages). We will show that this neuroscientific hypothesis is actually valid and that the prediction of intentions is a manageable, yet complex problem.

Actually, previous early activity recognition or prediction pipelines analyze motion patterns which are characteristic of an unfinished or future action, respectively, and such cues undoubtedly help to solve the task. For instance, to predict if two persons are going to give a high-five or shaking their hands (Vondrick et al. 2016; Lan et al. 2014), it is enough to detect a high/low wrist height during the first part of their interaction, respectively. Another important aspect of the current literature (e.g., see Kitani et al. 2012; Walker et al. 2014) is the use of the *context* to help the classification, namely, the objects present in the scene and the knowledge about the actions associated to them. Indeed, in real scenarios, context might not always be easily recognizable or may not contain enough information to discriminate among (similar but different) actions, being also misleading if the scenario is too noisy or cluttered (Stapel et al. 2012). However, while context is surely an important information for recognition, our test-case deliberately does not consider such clues and does not exploit such information.

Here, differently, grounding from the assumption that *the same class of motor acts can be performed with different intents* (Kilner 2011), we focus on exclusively analyzing such movements to figure out whether they actually embed a specific intention from the very beginning. Hence, we can face our problem by capturing the sole motion patterns which anticipate the intention while purposely excluding contextual information. Further, since the same brain areas are used in both motor planning and intent understanding (Oztop et al. 2005), the fascinating possibility of predicting intentions from the *kinematics only* is envisioned.

To prove our claims, instead of adapting existing benchmarks, we exploit a dataset specifically designed for intention prediction. Seventeen subjects execute several graspings of a bottle placed on a table, in order to either (1) pour some water into a glass, (2) pass the bottle to a co-experimenter, (3) drink from it, or (4) place the bottle into a nearby box. We acquire a dataset recording temporal 3D trajectories of 20 VICON markers outfitted on the subjects' hand and optical RGB videos from a lateral viewpoint, so that nothing but the table, the bottle, the subject's arm and a wall are visible (see Fig. 2). VICON and video data starts with the hand in a common resting position (for either Pouring, Passing, Drinking or Placing) and ends at exactly when the bottle is grasped. Hence, anything happening afterwards can be processed by our system, and no bias is introduced by leaving free the subject's hand initial position: we analyze grasping-a-bottle actions and, from sole these movements, we want to predict whether the bottle was grasped to pour, to pass, to drink or to place.

Another aspect which should be considered in the design of methods coping with action recognition problems and related variants is the capability of *generalization*. This results an even more crucial point for intention prediction as well. Specifically, since the same class of anticipative motor acts subsuming different intentions is executed by several subjects, not only we have to figure out intention-specific discriminants from similar motor acts, but such discriminants should be also transversal (i.e., invariant) across different subjects. Actually, we realize that a certain bias is also associated to the subjects executing the grasping actions, and we

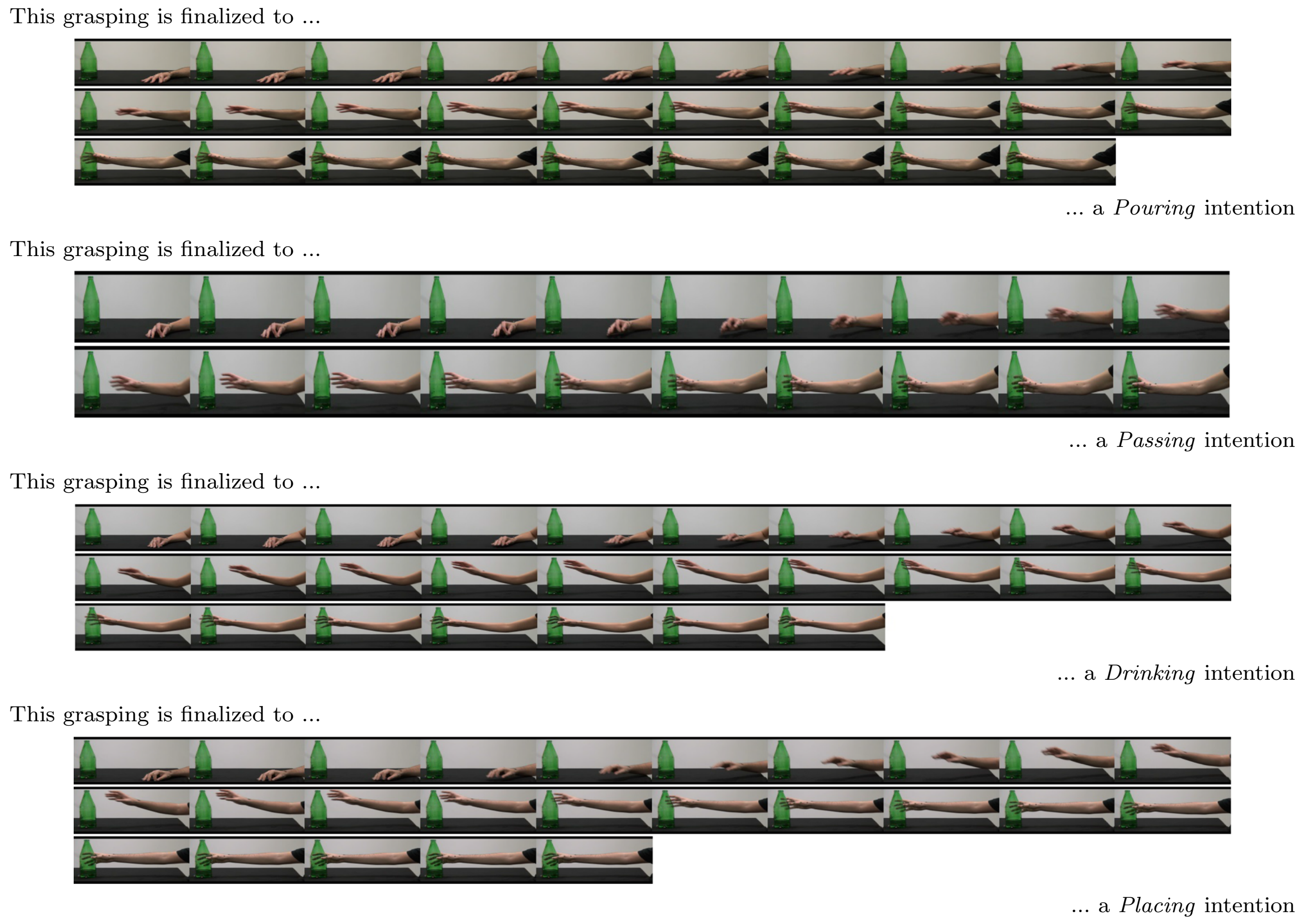

This grasping is finalized to ...
... a *Pouring* intention
This grasping is finalized to ...
... a *Passing* intention
This grasping is finalized to ...
... a *Drinking* intention
This grasping is finalized to ...
... a *Placing* intention

Fig. 2 The proposed problem of intention prediction. By only inspecting an apparently unrelated grasping-a-bottle motor act, we want to infer whether the latter is finalized to (1) pour some water into a glass, (2) pass the bottle, (3) drink from it or (4) place the bottle in a box (from top to bottom). We face this problem in pure kinematic terms: the context has been totally marginalized out. All sequences are ordered from left to right, top to bottom

tried to exploit such information in a novel way to increase the generalization power of our method. In order to cope with this additional complexity in an effective way, we propose an original approach derived from the domain adaptation research which considers each subject as a domain and adopt a novel subject-adversarial training pipeline to generalize better among the subjects. Our proposed approach showed the best performance in our test case, and also promising results in classic action recognition frameworks.

1.1 Main Contributions

To sum up, in this paper, we introduce **Intention from Motion** (IfM), a new problem of predicting human intentions (Pouring, Passing, Drinking, Placing), which all originate from a same, apparently unrelated motor act (grasping-a-bottle). We tackle this problem using the minimal possible information, that is, processing the kinematics only of the onset action, and deliberately not exploiting any contextual information: these are the necessary conditions to empirically prove the neuroscientific claim dictated by the problem above.

To this end, we develop computational methods to be applied to a dataset of grasping acts specifically designed to investigate intention prediction in which 17 subjects performed several grasping-a-bottle movements finalized to four different subsequent intentions (Pouring, Passing, Drinking, and Placing)[2].

We cast this problem in a *classification* scenario and we show that, despite IfM is arguably a challenging task, intention-specific discriminants are actually implicitly embedded in the grasping acts, and can be exploited by computational methods.

As most original contribution, we propose a method that explicitly addresses the biases associated to the human subjects performing the initial grasping action, an issue particularly affecting the intention prediction problem, even more severely than other action recognition paradigms. In particular, we discovered an inherent inter-subject variability and intra-subject similarity of the motor acts when performed by different and same subject(s), respectively, and we devised a method aimed at exploiting such information to improve its generalization ability, which is derived from the domain adaptation (DA) research (Csurka 2017). This is done by interpreting each training subject as a source domain and the unknown testing subject as target domain, being his/her intention labels *never* used in training.

The proposed method is named Subject-Adversarial Domain Adaptation (SADA) and it is formulated as an *unsupervised domain adaptation* problem (Ganin et al. 2016), where unannotated testing trials are used to promote both intention discrimination and subjects' confusion. As a generalization of SADA, we also consider the case where the testing trials are never exploited at all: the adaptation is in this case performed in a complete blind manner between all the training subjects only (i.e., the trials of the testing subject are never processed by the system in any way during training). This latter method, called Blind-SADA, can be interpreted as an unsupervised *domain generalization*

Some parts of this paper have been presented in recent conference works Zunino et al. (2017b), and Zunino et al. (2017a), mainly concerning the presentation of the intention prediction problem together with the processing (Zunino et al. 2017b) and the fusion of 3D and 2D data (Zunino et al. 2017a), respectively. This version extends both by thoroughly investigating the generalization aspect and proposing a new approach based on DA.

Paper Outline In Sect. 2, we discuss some related works from the literature. Section 3 introduces our dataset and provides a human recognition baseline. In Sect. 4, we present our subject-adversarial approach to solve the intention prediction problem, and experimental results are provided in Sect. 5. Section 6 reports some insights on 2D and 3D discriminants to classify the intentions. Finally, Sect. 7 draws the conclusions and sketches the future work.

2 Related Work

Early Activity Recognition (EAR) Ryoo (2011) first proposed a variation of bag-of-feature model to infer the ongoing activity by only analysing its beginning. The same problem was faced in Cao et al. (2013) with sparse coding. Hoai et al. (2012) designed a max-margin event detectors to early recognize emotions. Ryoo et al. (2015) proposed a dataset for early activity recognition from egocentric videos. Some works have attempted to investigate how much of the whole action is necessary to perform EAR by either a generative model (Davis and Tyagi 2006) or metric learning (Schindler and Gool 2008). Soran et al. (2015) devised a notification system for daily activities where, for instance, the detection of an ongoing milk boiling alerts the human user. Xu et al. (2015) formulates EAR as Internet-queries autocompletion. EAR is also tackled by Soomro et al. (2016) and by Kong and Fu (2016) by modifying the SVM framework, whereas, Ma et al. (2016) exploited Long Short Term Memory (LSTM) model on top of convolutional neural nets (CNN) architectures.

Action Prediction Li et al. (2012) used a random tree to model all the kinematics up to a certain instant and to predict the most likely future event (*e.g.*, predicting "grab an object" if "reach an object" is detected). Huang and Kitani (2014) proposed a pose-based approach for human interaction pre-

[2] Please, visit the official project page https://pavis.iit.it/projects/human-intention-prediction.

diction. Lan et al. (2014) developed the so-called *hierarchical movemes* to model human actions at multiple levels of granularities. Vondrick et al. (2016) and Kong et al. (2017) used the full past-future actions as data augmentation to train deep nets which can predict the future from the past. Jain et al. (2016) combined Recurrent Neural Networks (RNNs) and spatio-temporal features for action prediction, while encoder-decoder architectures were proposed in Bütepage et al. (2017) (multi-part network to process skeletal joints at multiple time-scales), and in Lu et al. (2017), the latter proposing an extrapolation model to generate the dynamics. Fermüller et al. (2017) predicted manipulating actions through LSTM networks fed with either accelerometer or video data.

Many frameworks exploited topic/probabilistic models to predict future actions by either modeling object-object/object-person relationships (Chakraborty and Roy-Chowdhury 2014; Li and Fu 2014) or detecting which areas must be used/avoided by vehicles/pedestrians during navigation (Kitani et al. 2012; Walker et al. 2014).

(Adversarial) Domain Adaptation Domain adaptation refers to the fundamental problem of *domain shift* (Daumé and Marcu 2006) between a *source* and a *target* dataset, used for training and testing, respectively. While the source dataset is fully annotated, the target dataset is not or can have only a few annotated samples. Consequently, either *unsupervised* and *supervised* adaptation pipelines have been proposed, depending on whether the labels in the target domain are used in training or not. Hand-crafted approaches have been proposed to learn transformations in order to align source and target domain through either dictionary learning (Huang et al. 2013; Shekhar et al. 2013), manifold projections (Gopalan and Li 2011; Gong et al. 2012) and covariance statistics (Sun et al. 2016; Fernando et al. 2013; Morerio et al. 2018; Volpi et al. 2018).

Concurrently to the recent deep learning revolution, many different architectures have been proposed to adapt between domains: encoders (Chopra et al. 2013), convolutional networks with either modified loss (Tzeng et al. 2014; Sun and Saenko 2016) or Maximum Mean Discrepancy to promote weights' sharing (Ghifary et al. 2014). Adversarial training has been proposed to *ad-hoc* techniques (e.g., the usage of soft labels (Tzeng et al. 2015) and gradient reversal layer Ganin et al. 2016) to perform a joint learning stage to devise a representation which is effective for the main classification tasks (object recognition Tzeng et al. 2015; Ganin et al. 2016 and re-identification Ganin et al. 2016) and, at the same time, invariant while shifting from the source to target domain. Regarding unsupervised domain adaptation, generative adversarial networks are well established: in Liu and Tuzel (2016), the joint distribution between the two domains is learned, Taigman et al. (2017) directly learns a transformation from the source to the target distributions. Instead, Tzeng et al. (2017) manages to directly transfer an end-to-end classifier trained on the source domain in a discriminative manner.

Novelty Aspects Differently to early activity recognition (Ryoo 2011; Hoai et al. 2012; Cao et al. 2013; Ryoo et al. 2015; Davis and Tyagi 2006; Schindler and Gool 2008; Soran et al. 2015; Xu et al. 2015; Soomro et al. 2016; Ma et al. 2016; Kong and Fu 2016; Cavazza et al. 2017a) and action prediction (Li et al. 2012; Huang and Kitani 2014; Lan et al. 2014; Vondrick et al. 2016; Jain et al. 2016; Kong et al. 2017; Bütepage et al. 2017; Lu et al. 2017), we do not observe the initial action we want to classify nor predict future actions from different past activities. Instead, we predict which intention (Pouring, Passing, Drinking, Placing) originates from the same class of motor acts (grasping-a-bottle). Our context-free setting is novel as opposed to a massive usage of contextual information (Chakraborty and Roy-Chowdhury 2014; Li and Fu 2014; Kitani et al. 2012; Walker et al. 2014). Also, the fixed starting position of the hand at the beginning of each trial is meant to remove any bias in the kinematics.

To the best of our knowledge, this is the first work which applies domain adaptation to action recognition and its variants. To this aim, we propose a domain adversarial adaptation method for intention prediction which learns from multiple source domains (i.e., multiple training subjects) how to adapt to an unknown target domain (i.e., new, never observed, testing subject). Ultimately, inspired by adversarial domain adaptation (Liu and Tuzel 2016; Taigman et al. 2017; Tzeng et al. 2017, 2015) we take advantage of multiple source domains in order to better generalize towards a target domain, even in the challenging case when the latter is not observed at all.

3 The Dataset

We utilized the dataset collected by the C'MON department, IIT. The dataset was designed as follows. Seventeen naive volunteers were seated beside a 110 × 100 cm table resting on it elbow, wrist and hand inside a fixed tape-marked starting point. A glass bottle was positioned on the table at a distance of about 46 cm and participants were asked to grasp it in order to perform one of the following 4 different intentions.

1. **Pouring** some water into a small glass (diameter 5 cm; height 8.5 cm) positioned on the left side of the bottle, at 25 cm from it.
2. **Passing** the bottle to a co-experimenter seating opposite the table.
3. **Drinking** some water from the bottle.
4. **Placing** the bottle in a cardboard 17 × 17 × 12.5 box positioned on the same table, 25 cm further.

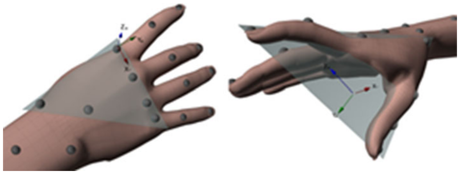


Fig. 3 The VICON marker geometry on the subject's hand

After a training session, the final dataset is composed by 253 trials of pouring, 262 of passing, 300 of drinking and 283 of placing - 1098 in total. For each, both video and 3D data have been collected. 3D marker trajectories and video sequences are acquired from the moment when the hand starts from a stable fixed position up to the reaching of the bottle, and both are precisely trimmed at the instant when the hand grasps the bottle, removing the following part. Our controlled setting constitutes an actual *worst-case scenario* since the context is fixed (so, not discriminative) for all the trials across subjects (*e.g.*, table size, location and size of the box, co-experimenter's position, etc.). Moreover, fixing the starting hand location and the bottle position, we put ourself in *neutral* conditions, removing possible subjective biases which might affect the classification performance (*e.g.*, some intentions might be better distinguished if starting hand position would have been left free).

3D kinematic data Near-infrared 100 Hz VICON system was adopted to track the hand kinematics. Nine cameras were placed in the experimental room and each participant's right hand was outfitted with 20 lightweight retro-reflective hemispheric markers (see Fig. 3). After data collection, each trial was individually inspected for correct marker identification and then run through a low-pass Butterworth filter with a 6 Hz cutoff.

Globally, each trial is acquired by means of a set of 3D absolute coordinates describing the trajectory covered by every single marker during execution phase. During the VICON's calibration stage, the floor plane is set to be the plane of the table and the origin of the reference system is defined over a fixed corner of the table, making it uniform across participant. The x, y, z marker coordinates only consider the reach-to-grasp act, the following movement is totally discarded. Indeed, the acquisition of each trial is automatically ruled by a thresholding of the wrist velocity $v(t)$ at time t, acquired by the corresponding marker. Being $\varepsilon = 20$ mm/s, at the first instant t_0 when $v(t_0) > \varepsilon$, the acquisition starts and it is stopped at time t_f, when the wrist velocity $v(t_f) < \varepsilon$.

2D Video Sequences Motor acts were also filmed from a lateral viewpoint using a fixed digital video camera (Sony Handycam 3-D) placed at about 120 cm from hand start position. The view angle is directed perpendicularly to the agent's midline, in order to ensure that the hand and the bottle were fully visible from the beginning up to the end of the movement. It is worth noting that the video camera was positioned in a way that neither the box (Placing), nor the glass (Pouring), nor the co-experimenter (Passing) were visible. Adobe Premiere Pro CS6 was used to edit the video in .mp4 format with disabled audio, 25 fps and 1280 × 800 pixel resolution. In order to format video sequences in an identical way to 3D data, each video clip was cut off at the exact moment when the bottle is grasped, discarding everything happening afterwards. To better understand how demanding the task is, note that the actual acquired video sequences encoding the grasping last for about one fourth of the future action we want to predict (see Fig. 2). Consequently all the sequences are about 20-30 frames long.

As evaluation procedure, we adopted the *one-subject-out* testing procedure, that is, we compute seventeen accuracies, training our system on all the subjects except the one we are testing and then we average all the accuracies to get the final classification performance. This testing procedure is more challenging than the usual cross-validation whose classification scores are always higher (see the Supplementary Material). We deem the adopted procedure is the correct one to devise a system, effectively able to better generalize and predict intentions in real world scenarios.

Another consideration is worth to be finally raised. When compared to other existing action recognition datasets, the controlled experimental conditions with which our dataset was built might seem a limitation. For instance, MPII-CAD (Rohrbach et al. 2012) and Salad 50 (Stein and McKenna 2013) cover more articulated (cooking) actions, while UCF-101 (Soomro et al. 2012) and HDMB51 (Kuehne et al. 2011) collect YouTube videos, thus guaranteeing a broad variability of backgrounds and context. Conversely, we deliberately designed our case study in order to properly answer the question whether the kinematics of a same ongoing action is enough informative to disclose the intention which will cause the following action. Indeed, the uncontrolled and real-world scenarios of the YouTube videos (such as in UCF-101 and HDMB51) may incidentally enrich the context with some cues that actually facilitate the prediction. Moreover, different future actions frequently begin with a quite different onset, e.g., two persons *approach* each other before a "kissing" action occurs, or people *rise their hands* before a "high-five" action is carried out (Lan et al. 2014). Additionally, in some cases (MPII-CAD and Salad 50, for instance), the prediction is facilitated by the detection of *which* object (out of many others) is grasped (e.g., a knife to predict "cutting").

In our case instead, we want to predict *why* the same object (bottle) is grasped, therefore complicating the applicability of existing prediction pipelines to our problem (Sect. 5.4). In such scenario, existing state-of-the-art techniques, although effective to some extent, are not so well performing.

3.1 Human Performance in IfM

Predicting intention from motion is not a trivial task for human observers. Studies investigating the ability of human observers to discriminate intention from motion using two-choice decision tasks report accuracies in the range of 49–68% (Cavallo et al. 2016; Koul et al. 2019). For review, see Becchio et al. (2018).

As a preliminary analysis to check how human beings can predict intentions, we tested the human capabilities on Pouring versus Placing and Pouring versus Drinking throughout the following experimental apparatus. We asked each of 18 participants to watch 400 videos of reach-to-grasp movements and predict whether it was finalized either to pour some water or pass the bottle. We balanced the videos from each class (50 for Pouring and 50 for Placing, with 4 repetitions of each video). The experiment starts showing the complete execution setup of the reach-to-grasp and its conclusion (Pouring or Placing) in a wide zoom where the glass or the box, respectively, were visible. Then, we narrow the field of view, discarding everything except the arm, the table and the bottle, and we show only the reach-to-grasp movement, exactly as the videos processed by the algorithms. After 8 demo trials in which the future intention was revealed, we randomly shuffled the 400 videos and tested all the participants, registering their guess. Averaging all the human accuracies in the Pouring versus Placing test, we get 68% of accuracy. Afterward, we move to the second test (Pouring versus Drinking) and we repeated the same procedure for a different set of 18 participants. In Pouring versus Drinking, accuracy decreases to 58% (−10%).

Thus, although there are cases where the human brain can read in a grasping act some motion pattern which anticipates its intention, we can anticipate that computer vision methods will prove to be more valuable outperforming human predictive ability (Table 1).

4 Subject-Adversarial Adaptation for Intention Prediction

The capacity of generalization is indeed a requested ability of any (action) recognition method. This is even more important in our case since the actual intention is never observed and such discriminants should be spot from very similar grasping actions. Indeed, in IfM, a better generalization can be implicitly achieved by identifying intention-specific subject-invariant discriminants which are embedded in the kinematics of a generic grasping motor act. To cope with this problem, in this Section we propose a new approach able to *explicitly* promote subject-independence for predicting intentions. This allows to actually improve cross-subject generalization and, consequently, prediction performance.

Table 1 Human recognition performance

	Human performance (%)
Pouring versus placing	68
Pouring versus drinking	58

Specifically, in Sect. 4.1, we will present a novel approach to action recognition which is able to exploit these biases to improve the generalization capacity of the method, resulting in an ultimate superior performance for intention prediction. In Sect. 4.2 we will provide the details for reproducibility purposes.

4.1 Subject-Adversarial Domain Adaptation

To reduce the bias generated by the different agents, we resort to the idea to explicitly consider such information in devising a training method able to “confuse” the subjects such as to increase the generalization ability of the classification model. To this end, we propose a novel approach to action recognition which is based on a well established unsupervised domain adaptation technique (Ganin and Lempitsky 2015). The main goal of unsupervised domain adaptation techniques is to perform well not only on samples drawn from the training data distribution (source), but also on samples drawn from other distributions (target), whose labels are not known at training time.

Leveraging on the subjects’ related biases discussed above, we pursue the original perspective of combining action recognition (and variants) with domain adaptation. By identifying each subject as different domain, we subsequently propose to perform adaptation in a multi-domain case where we exploit multiple (source) subjects in order to adapt our models to perform well on a new, unknown (target) agent.

We adopt adversarial domain adaptation for action recognition by learning a shared feature representation between subjects which can be effective for intention disambiguation. We want to learn a representation which, at the same time, leads to a top-scoring intention classifier and to a random chance scoring discriminator of the subjects (Ganin et al. 2016). This can be obtained employing adversarial training by means of the min-max formulation described in the following.

We consider each grasping a bottle movement in our dataset $\mathcal{D}$ as a triplet $[\mathbf{x}, s, y]$, where $\mathbf{x}$ is an arbitrary feature vector encoding it, s is the subject’s label, and y is the intention’s label (see Fig. 4).

We look for a feature representation $\mathbf{f}(\mathbf{x}|\mathbf{W_f})$, depending on some parameters $\mathbf{W_f}$, which is trained to be intention-

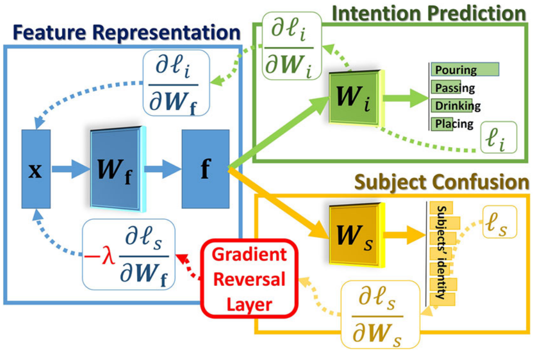


Fig. 4 The adopted Subject-Adversarial Neural Network (SANN) (Color figure online)

discriminative and subject-invariant. This is achieved through the following optimization problem

$$\min_{\mathbf{W_f}, \mathbf{W}_i} \sum_{[\mathbf{x},s,y]\in\mathcal{D}} \ell_i(y, g(\mathbf{f}(\mathbf{x}|\mathbf{W_f}), \mathbf{W}_i)) \tag{1}$$

$$\max_{\mathbf{W}_s} \min_{\mathbf{W_f}} \sum_{[\mathbf{x},s,y]\in\mathcal{D}} \ell_s(s, h(\mathbf{f}(\mathbf{x}|\mathbf{W_f}), \mathbf{W}_s)). \tag{2}$$

on top of the feature representation $\mathbf{f}$. Precisely, Eq. (1) promotes an accurate prediction of intentions: the loss function ℓ_i is minimized as to penalize discrepancies between the actual intention label y and the high-level embedding g which is trained to be discriminative for the sake of the intention prediction task. In (2), we still consider a similar setup in which we train a high-level encoding h by mean of a loss function ℓ_s which consider the subjects' identity s. This second loss function is minimized with respect to the weights $\mathbf{W}_f$ which defines the feature encoding $\mathbf{f}$, being at the same time maximized at the classifier level—that is, the weights $\mathbf{W}_s$. The whole idea is to deploy an adversarial game in which we want to train at our best an effective feature encoding $\mathbf{f}$ which is effective in predicting intentions, without suffering of the retrieved subjects' related biases. Concretely this is achieved by a multi-task network (intention prediction and subject confusion), where the two heads of the architecture are intentions and subjects classifier, respectively. We try to get rid of the subjects' biases by achieving a random chance classifier for subject identities: this will implement the idea of having feature representations which are totally invariant across different subjects. At the same time however, we pretend the very same feature representation to be discriminative enough to allow a reliable prediction of intentions (and therefore we train the intention prediction branch to be as effective as possible).

As far as we know, our work is the very first attempt of applying (adversarial) domain adaptation to action recognition problems (including the variants presented in Fig. 1). In this work, we demonstrate that this class of methods is indeed suitable for intention prediction by considering the following two settings.

Subject-Adversarial Domain Adaptation (SADA) The SADA approach is derived from the unsupervised domain adaptation pipeline (Ganin and Lempitsky 2015) where the un-annotated target data (here, the testing subject) is used to modify the feature representation, while the learning phase of the classifier for the main task is done on the source domains only (here, the training subjects' actions). In practice, the source domain data is used to learn the classifier to discriminate actions (intentions in our case) in a supervised way, whereas the target domain data is still used in training, but in an unsupervised way, since action labels are unknown (we only use the information that the test subject's identity is different from that of any other training subjects). In our case, the actions of the test subject are our target domain while the actions of all the other subjects constitute the source domain: we aim at training the system by improving the action classification performance while minimizing the capability of the system to identify the subject who executed that action.

Blind-SADA Blind-SADA can be seen as a generalization of the classical domain adaptation setting that, overall, relies on the fact that the target domain is fixed and specified. In fact, even in the unsupervised case, un-annotated target data are exploited during learning to adapt with the source. Here, differently, we posit that the availability of multiple source domains (i.e., training subjects) can provide enough information as to learn an adaptation which is enough powerful to be *blindly* applied to an arbitrary target domain (i.e., testing subjects), without exploiting target data in any way during the learning stage (*domain generalization*). We also explored this setting since in a general video-surveillance framework, a system should be able to perform well on a variety of unknown, never seen, testing subjects, still ensuring a high generalization in predicting humans' intention. In this situation, it is desirable to investigate whether subjects' confusion (2) applied on a fixed number of subjects is still generalizable to other, unseen ones. In our experiments, we do this by optimizing Eqs. (1) and (2) by only using the data of 16 subjects, without using the data of the test subject left out neither for subjects' confusion (differently from SADA), nor for tuning the parameter of the intention prediction branch (1) (see Fig. 4).

4.2 Implementation Details

Technically, SADA and its Blind variant are implemented by the architecture inspired by Ganin et al. (2016) and named *Subject-Adversarial Neural Network* (SANN), which is composed by 3 modules. A first, low-level, network module learns the feature representation $\mathbf{f}(\mathbf{x}|\mathbf{W_f})$—blue in Fig. 4. After that, two separate intention—(green in Fig. 4) and subject-related modules (yellow in Fig. 4) are responsible for the intention-discrimination and subjects-confusion, respectively. As previously explained (Sect. 4.1), the blue module is responsible for achieving a representation which, at the same time optimizes the green module (intentions) and fool the yellow one (subjects). Therefore, our adversarial approach stands from the fact that the yellow module seeks for a perfect subjects' discrimination built on top of a feature representation which is learnt to be subject-invariant in the blue module.

SANN is trained accordingly to the one-subject-out protocol. It is important to note that, for all-class comparison considered, we train one SANN per subject left out for testing, using the remaining subjects as multiple source domains. Performance of each network are evaluated on the subject left out and results are averaged across.

More specifically, in SADA, the subject confusion module is fed with the unlabeled (as for the intention) data of the testing subjects to adapt the feature $\mathbf{f}(\mathbf{x}|\mathbf{W_f})$ to the specific agent. Differently, Blind-SADA *never* exploits the trials of the testing subject in training and performs adaptation by totally ignoring the target domain (both identities and intention labels).

We accommodate the publicly available code[3] of Ganin et al. (2016) to deal with a different number of subjects to perform adaptation. Indeed, Ganin et al. (2016) considers a simplified setting of one target domain only, whereas, differently, we consider multiple domains. The optimization of (1) and (2) is carried out by using a joint back-propagation In particular, we compute the updates on the parameters $\mathbf{W}_s$ and $\mathbf{W}_i$ separately on the two branches. Then, we used the gradient reversal layer (Ganin et al. 2016) to change the sign of the derivative of the subject loss ℓ_s with respect to $\mathbf{W_f}$ (after a re-scaling by a parameter λ). The derivative of ℓ_i with respect to $\mathbf{W_f}$ is instead back-propagated with the correct sign (see Fig. 4). A multi-layer perceptron (MLP) network with one hidden layer of dimension 200 was designed as the shared feature representation $\mathbf{f}(\mathbf{x}|\mathbf{W_f})$. For the intention prediction module, we trained a four-way softmax function using a cross entropy loss for ℓ_i. Similarly, for the subject confusion module, a 17- or 16-way cross-entropy loss is used for ℓ_s in SADA and Blind-SADA, respectively.

We cross-validate λ by selecting the value which maximally fool the subjects' classifier in the subject confusion module.

5 Experiments

This Section reports an extensive experimental analysis of our work. We first setup and illustrate the baseline against which we will compare our proposed SADA approach. Specifically, Sect. 5.1 shortly presents the baseline methods applied to either 3D VICON markers or 2D video sequences and show the related results. In Sect. 5.2, we discuss the obtained performances with special focus on the challenges related to generalize intention prediction across different subjects. Then, we illustrate in Sect. 5.3 the results obtained by our proposed approaches SADA and Blind-SADA, and we discuss them, also in comparison with baseline scores and with respect to the existing compliant prediction pipelines in the literature (Sect. 5.4). Finally, as a collateral experiment, in Sect. 5.5 the proposed subject-adversarial adaptation approaches are evaluated with respect to similar setups for action recognition.

5.1 Baseline 3D and 2D Methods

Several 3D and 2D action recognition techniques are proposed in the literature, and a complete review of them is out of scope of this paper, also because the majority of these methods does not fully fit with our scenario.

State-of-the-art approaches for processing 3D skeletal joints either consider deep learning applied to encode and classify raw data or hand-crafted kernel representations—which codify human motion—followed by classification (Moeslund et al. 2006). As shown in Cavazza et al (2017b), kernel methods are undoubtedly the most performing methods taking into consideration that we are not on a big data regime. Also, deep neural networks have so far been applied to the classification of coarse actions (such as running or clapping), but not yet to the prediction of fine motor acts (such as grasping-to-pour versus grasping-to-pass). Therefore, in this paper we have only considered kernel methods and within this family of methods, we exploited a covariance-based representations following the state-of-the-art approach of Cavazza et al. (2016) in which a kernel matrix (computed out of un-normalized x, y, z coordinates) is combined with the usual covariance operator to devise a powerful representation which is further elaborated for max-margin classification.

Within the available approaches to handle video data (Moeslund et al. 2006), we have considered optical-flow based dense trajectory features (Wang et al. 2013). Alternatively, we take advantage of the well established capability of deep neural networks to produce frame-based representations: to do so, we extract fc7 features out of fine-tuned AlexNet fed with optical flow (OF) images, after a classical spatial resize to 227 × 227. We computed OF images with three channels constituted by the horizontal and vertical components, and the magnitude of the optical flow field, after a

[3] http://graal.ift.ulaval.ca/dann/.

Table 2 Selected feature encodings of 2D and 3D data for IfM

	ker-COV [3D] (%)	DT-HOF-*VLAD* [2D] (%)	AlexNet-OF-*VLAD* [2D] (%)	Human performance
Pouring versus placing	91.87	86.18	94.18	68%
Pouring versus drinking	91.58	81.48	77.95	58%
Pouring versus passing	81.69	74.44	74.20	–
Passing versus drinking	87.64	71.53	66.05	–
Passing versus placing	75.46	75.15	94.68	–
Drinking versus placing	91.24	79.23	96.18	–
All-class	73.72	58.23	65.64	–

We consider all the possible pairwise comparisons between intentions and the all-class one. For the sake of the comparison, we also report the results of human performance

preliminary normalization in the range [0, 255]. In order to pool dense trajectory/CNN features into a unique representation to encode each video, we used a VLAD encoding (Jégou et al. 2010), square-root normalized, followed by a linear support vector machine (SVM), with the cost parameter $C = 10$.

The performance of the aforementioned features is reported in Table 2. Note that such approaches were selected out of an extensive pool of 3D and 2D descriptors on the basis of the performance. Readers can refer to the Supplementary Material for additional details.

5.2 Discussion

The baseline methods are able to provide a relatively solid classification performance. This certifies the fact that grasping gestures already codify motion pattern relative and specific to the actual intention which originates the grasping itself. Although human performance is verified on a subset of videos, computational methods seem to outperform human observers. Therefore, the actual execution of the grasping is modified accordingly to the intention which needs to be fulfilled. Such variations are actually captured by means of the 3D and 2D encodings that we tried, ultimately certifying that, as a computer vision problem, predicting intention from motion is a feasible problem. Even when the context is uninformative and the unique source of information is provided by the kinematics. The accuracy results provided in Table 2 are averaged across all the 17 subjects available in the dataset. A preliminary quantitative analysis to assess the bias among the subjects can be estimated by calculating the standard deviation (std) related to the average all-class performance of the 3 best baseline methods: std values result 13.40%, 14.14% and 16.12%, for ker-COV, DT-HOF-*VLAD* and AlexNet-OF-*VLAD* features, respectively.

As one can note, the standard deviation values are pretty high, meaning that accuracies are largely variable among the subjects. In other words, the generalization (subject independence) reached by the models on the new testing subject is not so high, which gives margin for improvement.

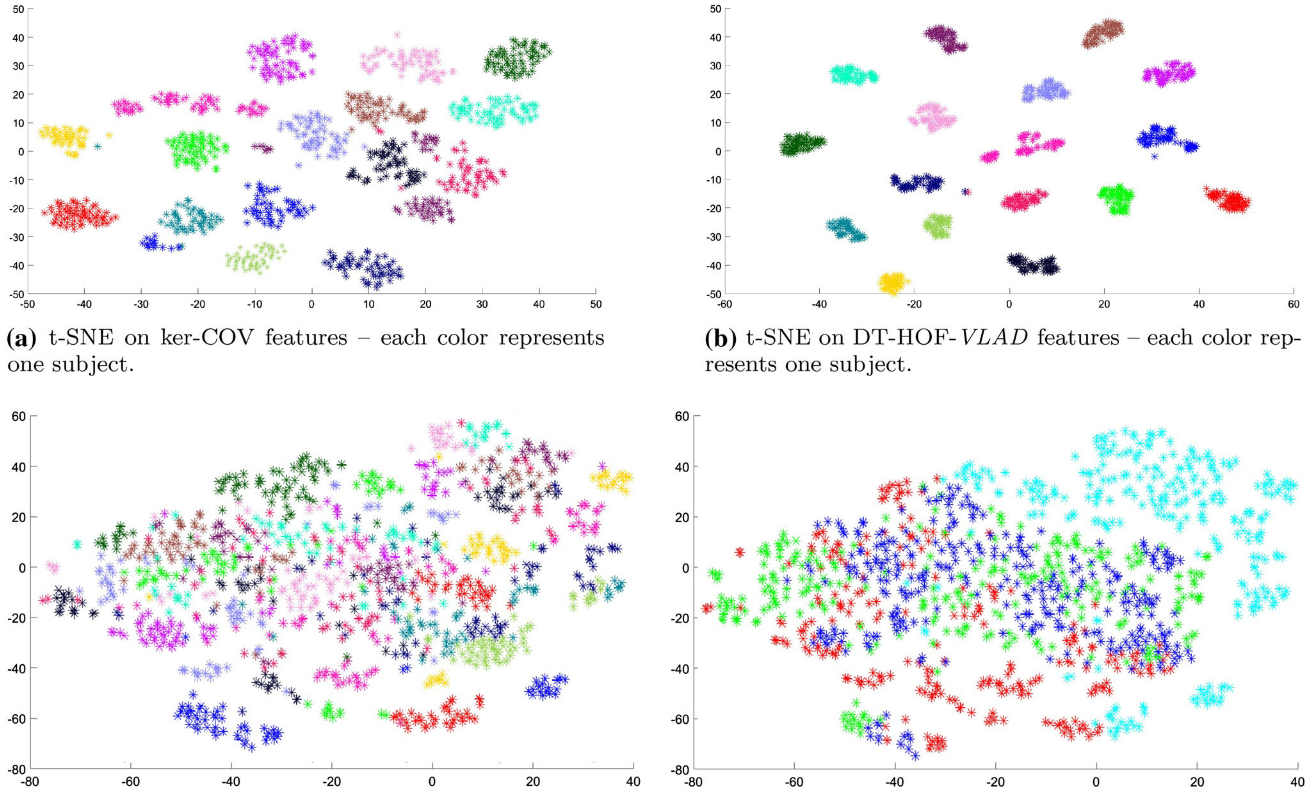
To further verify such claim, inspired by Zunino et al. (2017c), we performed another experiment in order to measure the bias provided by each subject. In Zunino et al. (2017c), leveraging quantitative evidence of the high variances associated to the same action performed by different subjects, action recognition is formulated as a 2-stage pipeline where, 1) the subject is identified and 2), its actions are recognized. Interestingly, for the task of subjects' identification, the same features exploited for discriminating actions are used, further denoting a clear evidence of the subject-related bias.

Inspired by this idea, in our case we used the baseline features (Table 2) to train a multi-class SVM to identify the 17 subjects in the proposed dataset. To do so, we adopted a one-intention-out testing protocol where every trial referring to one single intention was left out for testing, while all the remaining trials were used for training. In Table 3, we report the subjects' identification performance obtained after averaging across each intention left out. We register an outstanding performance of both ker-COV and DT-HOF-*VLAD* for subjects' identification, suggesting that intention prediction has a much stronger subject related bias with respect to classical action recognition problem. Differently, the performance of AlexNet-OF-*VLAD* is lower: presumably, after fine-tuning the network, a good intention prediction performance is achieved by implicitly bridging the subject-related biases.

The results in Table 3 are also corroborated by the t-distributed Stochastic Neighbor Embedding (t-SNE) technique (Van der Maaten and Hinton 2008), the most used state-of-the-art data visualization method. We applied it to ker-COV, DT-HOF-*VLAD* and AlexNet-OF-*VLAD* features, obtaining the plots in Fig. 5. Let us stress that t-SNE is a fully unsupervised method which does not exploit neither actions' nor intentions' labels. Nevertheless, ker-COV and DT-HOF-*VLAD* representations are perfectly able to cluster in 17 groups, each one corresponding to a single subject. The information of the subject who performed the grasping is clearly present in such representations and this can be seen

Table 3 Subjects' identification performance

ker-COV [3D]	DT-HOF-*VLAD* [2D] (%)	AlexNet-OF-*VLAD* [2D] (%)
97.25%	100	53.34



(a) t-SNE on ker-COV features – each color represents one subject.

(b) t-SNE on DT-HOF-$VLAD$ features – each color represents one subject.

(c) t-SNE on AlexNet-OF-$VLAD$ features – each color represents one subject.

(d) t-SNE on AlexNet-OF-$VLAD$ features – each color represents one intention.

Fig. 5 Bi-dimensional embedding of ker-COV, DT-HOF-*VLAD* and AlexNet-OF-*VLAD* using t-distributed Stochastic Neighbor Embedding (Color figure online)

as a bias that needs to be removed when training an intention predictor (see Fig. 5a, b). On the other hand, we are also able to explain why AlexNet-OF-*VLAD* features are not suitable in classifying the subject (see Fig. 5c). In fact, the t-SNE plot (in Fig. 5d) shows how, apparently, the fine-tuning process has achieved a nice separation of Placing intention (in cyan) versus the others by mixing all the subjects.

In summary, we have empirically proved the existence and the impact of subject-related biases for intention prediction, being this trend more critical than in action recognition (Zunino et al. 2017c). Therefore, achieving intention prediction in a generalizable manner across subjects is a difficult task, being nevertheless paramount for deploying an actual recognition system. In the following section, we will discuss the results obtained by SADA approach which properly tackle this problem of generalizing across subjects and show that reducing this bias is beneficial for the sake of predicting intentions.

5.3 SADA and Blind-SADA Performances

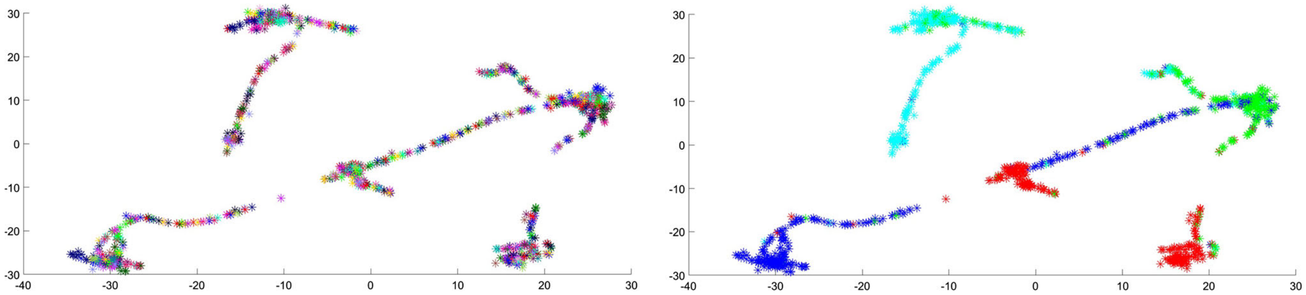
In Table 4, we report the results corresponding to SADA and Blind-SADA, as compared with a baseline methods of Table 2. The SANN architecture, takes as input the selected features (**x** in Fig. 4) ker-COV, DT-HOF-*VLAD* or AlexNet-OF-*VLAD* as in Table 2.

As a baseline approach to compare our subject-adversarial domain adaptation, we run an experiment where the yellow branch of Fig. 4 (subject confusion) is not present and we only consider the blue (feature representation) and green branches (intention prediction). Technically, one can frame the baseline as training a multi-layered perceptron (MLP) applied to intention prediction and fed with all the selected features from Table 2. As a common pre-processing step on data, we run PCA on the ker-COV, DT-HOT-*VLAD* and AlexNet-OF-*VLAD*, retaining the 99.5% of explained variance: this step is only required to speed up the computation and we did not register a major effect on performance. The achieved

Table 4 Subject adaptation results on IfM

	Baseline	Blind-SADA	SADA
ker-COV [3D]	71.57	73.13 (λ = 0.6)	**80.48** (λ = 0.1)
DT-HOF-*VLAD* [2D]	56.01	57.15 (λ = 1.5)	**70.42** (λ = 0.2)
AlexNet-OF-*VLAD* [2D]	65.64	66.59 (λ = 1)	**67.95** (λ = 0.1)

In brackets the best setting of λ
Best accuracy values are highlighted in bold



(a) t-SNE using ker-COV features and Blind-SADA – each color represents one subject.

(b) t-SNE using ker-COV features and Blind-SADA – each color represents one intention.

Fig. 6 Bi-dimensional embedding using t-Distributed Stochastic Neighbor Embedding for the ker-COV features transformed with Blind-SADA (Color figure online)

all-class performance is comparable with the one of obtained with SVM classification in Table 2.

Instead, in Table 4, we note a large improvement using SADA in the main intention prediction task: +8.91% for ker-COV, +14.41% for DT-HOF-*VLAD*, and +2.31% for AlexNet-OF-*VLAD*. The largest improvements are obtained considering DT-HOF-*VLAD* for video data and the 3D-based ker-COV encoding. These two neatly increased scores support the t-SNE plot in Fig. 5a, b, showing almost perfect compact clusters per subject with ker-COV or DT-HOF-*VLAD* features. This framework proved to be able to remove the predominant subject bias information from the data samples and to get better performance in the multi-class intention prediction task.

The CNN features deserve a separate discussion since the improvement with domain adversarial training is not huge although still present. We guess that the fine tuning process operated for CNN feature extraction already reduces the impact of the subject-related biases to some extent. In other words, CNN fine tuning already performs a sort of domain adaptation and subject confusion (as visible in t-SNE plots in Fig. 5 and 5), hence our framework is less effective in this case.

The results of Blind-SADA are reported in the second column of Table 4. The improvement with respect to the baseline approach is smaller than SADA, but still significant: +1.56% for ker-COV, +1.14% for DT-HOF-*VLAD*, and +0.95% for AlexNet-OF-*VLAD*. Hence, we can still assert that training the net with the proposed SANN framework is effective for intention prediction. This means that, also relaxing the classic domain adaptation framework, subject confusion is also beneficial when the target domain is not utilized during training, since a hidden representations could still be learnt to discriminate better between the intentions, reducing the noisy knowledge (i.e., the bias) coming from the subject identities.

To get a deeper insight on how the features are transformed by means of Blind-SADA training process, we plot in Fig. 6 the hidden representation of ker-COV when one subject trials are left fully out (in this case, subject 1). If we compare the t-SNE representations in Fig. 5 with those in Fig. 6a, b, we can note that the new ker-COV hidden representations are no more grouped in compact clusters associated to subjects. In Fig. 6a, the subjects are totally mixed whereas the samples are rearranged better for the main intention prediction task, as visible in Fig. 6b. This suggests that the training process has still learned feature discriminants for the intentions, at the expense of making indistinguishable the subjects, which was exactly our goal. Actually, if we try now to perform the subjects' identification experiment over the hidden representation plotted in Fig. 6a, b, the average accuracy drops from 97.25% (Table 3) to 6.88% coherently obtaining an almost random chance performance in subject identification.

5.4 Evaluation of Existing Prediction Pipelines

We posit our prediction problem as a classification one, according to the evidence that the onset of an apparently

Table 5 Subject adaptation on MSR-Action3D and HDM-05 action recognition benchmarks using COV features

	Baseline	Blind-SADA	SADA
MSR-Action3D	80.41	81.73 ($\lambda = 0.1$)	**84.84** ($\lambda = 0.6$)
HDM-05	94.68	95.41 ($\lambda = 0.2$)	**95.93** ($\lambda = 0.4$)

In brackets, the value of λ adopted
Best accuracy values are highlighted in bold

unrelated action is sufficient to quite reliably guess the future intention. Indeed, this is true when we accurately analyze 3D data, and we actually noted that in general the 3D encoding shows a superior performance with respect to the 2D case, even when deep features are utilized. Besides, it is important to highlight that also humans can predict intentions to a certain extent (Cavallo et al. 2016), and they still do not use 3D data. However, it is difficult in these cases to decouple the contribution given from the kinematics from that given by the context. Nevertheless, recent literature proposed action prediction pipelines and we would like to test also these approaches – when hypotheses fit our dataset conditions – with the aim of gaining in performance and bridging the gap with the 3D case.

Despite the broad literature in early activity recognition and action prediction, most of approaches are not directly applicable to the IfM problem. Indeed, (Lan et al. 2014) relies on a fine decomposition of the activity into coarse, mid-level and fine actions classes: of course, this is not applicable to our simple grasping movements. In Vondrick et al. (2016), a CNN is trained by jointly considering the present and the future frames of a given scene, while, in our case, only the (present) graspings are exploitable as data. Despite Koppula and Saxena (2016) deals with grasping motor acts as we do, it only predicts *which* object is grasped, not *why*, as we aim at. Finally, Hoai et al. (2012) and Li et al. (2012) need massive annotations of the emotion disclosure and actionlets, respectively, while, in this sense, our problem is totally unsupervised.

Among the few works directly applicable to our problem, we evaluate the temporal tessellation and dynamic bag-of-word histograms proposed in Ryoo (2011). Using this algorithm, the all-class classification accuracy results 45.12%, which suffers a gap of − 10.89% and − 20.52% with respect to the baselines DT-HOF-*VLAD* and AlexNet-OF-*VLAD*, respectively. Note that the previous performances do not take into account the proposed subject-adversarial adaptation method which gives a further boost, increasing the gap of about 4 percentage points. Thus, globally, despite all the aforementioned prediction pipelines are really powerful in their experimental conditions, the same methods seem little effective in our setting in which subject-adversarial methods are more successful.

5.5 SADA on Action Recognition Datasets

Grounding on the experimental findings discovered in IfM, we aim now to assess SADA framework in public action recognition datasets to see if we can gain in accuracy performance. The necessary condition in which we can apply the proposed pipeline is having access to action and subjects labels at the same time for each trial. For this purpose, we considered the public MSR-Action3D (Li et al. 2010) and HDM-05 (Müller et al. 2007), using the off-the-shelf covariance feature (COV) representation utilized in Zunino et al. (2017c), and demonstrated to be the most effective representation for IfM.

Since we cast intention from motion problem as a classification task, it is fair to test the same method in a general action recognition pipeline which involves several subjects who perform different gesture classes. Please, note that the main focus of this Section is not to present state-of-the-art results on action recognition benchmarks, which are 97.4% and 99.0% for MSR-Action 3D (Cavazza et al. 2019) and HDM-05 (Xie et al. 2018), respectively. Rather, in considering such action recognition benchmarks, we simply aim at transferring the setup used in the intention prediction paradigm without any peculiar customization, with the idea of showing that our proposed subject-adaptation approach can be effective even using off-the-shelf representations.

As done for IfM, we carried out the one-subject-out testing procedure where each subject is left out for testing, and we fed COV features into the baseline, SADA and Blind-SADA architectures. As previously done, λ is chosen by cross-validation by selecting the value which best achieves subject confusion. The results in Table 5 show a trend similar to the figures shown in Table 4: performance improves passing from the baseline to the Blind-SADA, finally registering the largest improvement of +4.43% in MSR-Action3D and +1.25% in HDM-05 using the SADA framework. Therefore, on these datasets we retrieve the same findings of IfM, certifying that the same approach can also be beneficial for the classic action recognition problem.

6 Mining Discriminants for Intentions

Leveraging the classification performance obtained with the proposed subject-adversarial training methods (Sect. 5), we can state that the IfM is a complex, yet manageable problem. In fact, we found that it is possible to predict the intention with which a motor act is performed, even in absence of contextual cues. In spite of that, we may still wonder why such results were achievable, possibly finding some spatio-temporal cues in the grasping movement that are more informative than others. To this end, we exploited three different approaches. In Sect. 6.1, we provide some "visual interpretation" to ground

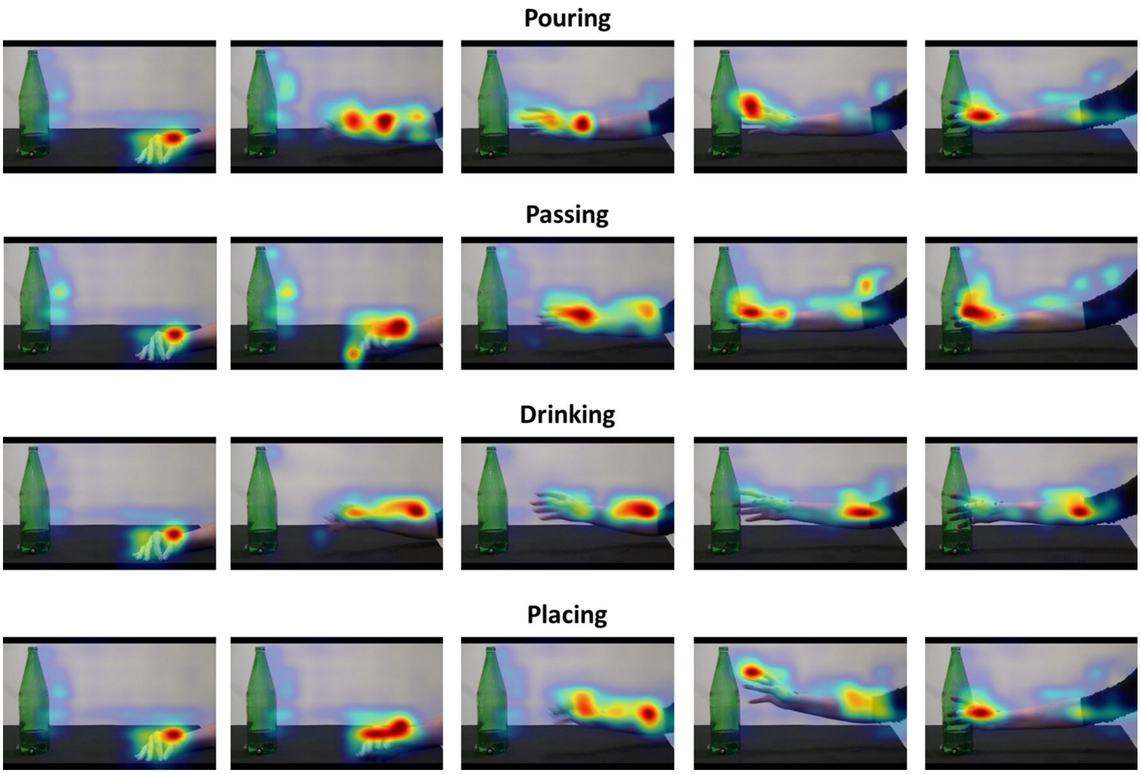


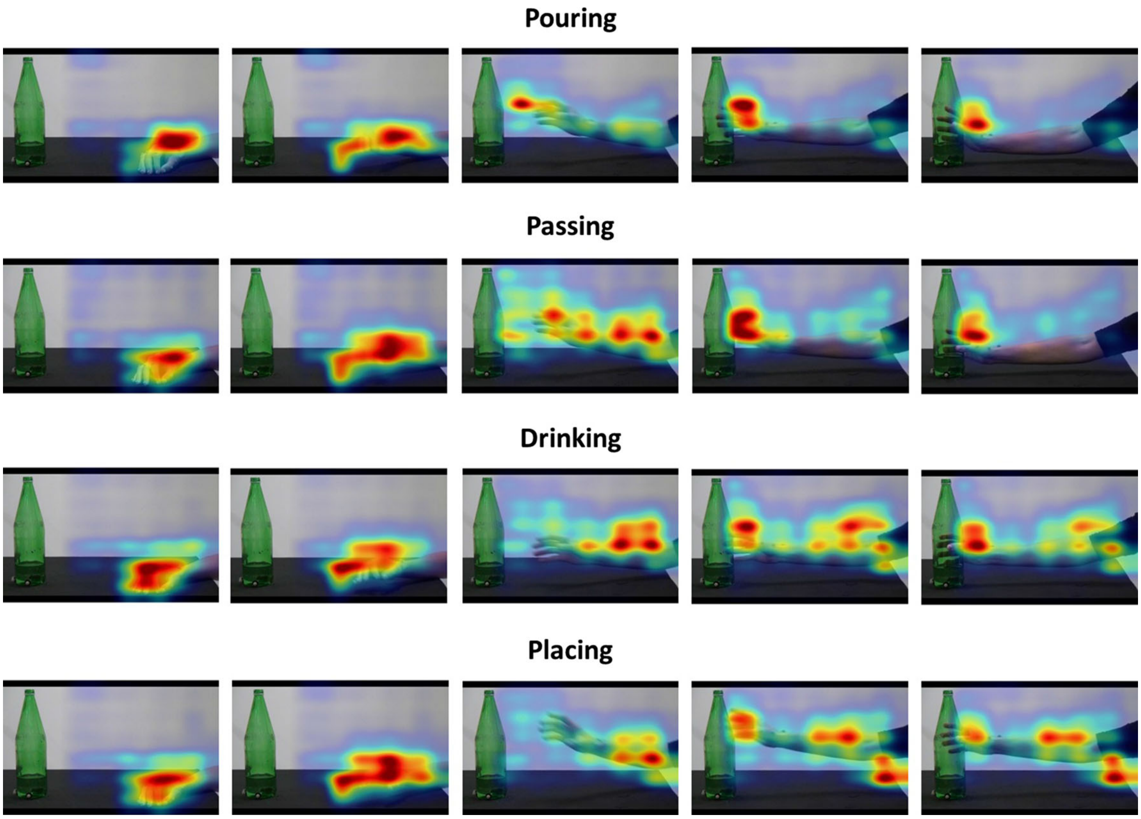
Fig. 7 Mining discriminants for intentions in video sequences. We show exemplar frames from correctly classified videos taken from a specific subject in IfM dataset, one for each intention. The method Bargal et al. (2018) always localizes the spatial-temporal discriminant cues around the hand and the arm of the subject. For this specific subject the saliency maps for the Drinking intention are highlighting mostly the elbow parts while mainly focusing on the wrist for the other ones

the evidences used to predict the intentions at the frame-level. In Sects. 6.2 and 6.3, we perform a similar analysis at the feature-level and at the level of raw 3D joints data, respectively.

6.1 Mining Discriminants for Intentions in Video Sequences

Nowadays, there is an increasing interest from the computer vision community in trying to "understand" what a deep learning model has learnt and which parts of the input data are more useful to support the prediction. All these approaches are mainly devised for image recognition and very few try to do the same for video understanding task. The recent work Bargal et al. (2018) is well suited for our purpose since it presents a top-down saliency approach able to localize spatial-temporal segments within a video that correspond with a specific action. Given a CNN+LSTM model trained for video action recognition, a standard forward pass is performed to activate the neurons. By means of an ad-hoc backward pass, one can then compute and propagate winning neuron probabilities normalized over space and time, following the excitatory connections inside the architecture down to the video frames. This process yields action-discriminative saliency maps, highlighting the most discriminative patches in the relevant frames within an input video. As to replicate the CNN+LSTM setup, we used VGG16 to extract `fc7` features from each video. In turn, these features are recursively given in input to a single layered LSTM (256 hidden units), which is jointly trained with the CNN for predicting the correct intention. We still considered the one-subject-out protocol, training one of such architectures per subject who was left out for testing. This procedure results in a 58% average test accuracy in the all-class comparison.

In Fig. 7, we report some visualizations of saliency maps overlapped on video frames corresponding to 4 correctly classified graspings actions (performed by Subject 10) for the 4 different intentions. We can notice that the highlighted parts (in red) belong to the subject's hand and arm mainly, and similar patterns occur when varying the subject. Please refer to the Supplementary Material for analogous saliency visualization of other subjects. Following the insights of Bargal et al. (2018), we have also computed the sum of the spatial saliency maps for each frame and we found that its distribution is almost homogeneous in time, meaning that there are not meaningful peaks (in time) which trivially helps to disambiguate intentions. This result is coherent with our snippet analysis (reported in the Supplementary Material): there



Pouring

Passing

Drinking

Placing

Fig. 8 Mining discriminants for intentions in video sequences. Using the method of Bargal et al. (2018) we visualize an integral saliency per intention, averaged over all the subjects and trials. Drinking and Placing intentions appear on average more discriminative in the elbow movement while for Pouring and Passing the saliency map is focused on the wrist movement

exists no specific temporal instant of the grasping action which is trivially more useful than others when predicting intentions.

Finally, we carried out an integral analysis of saliency across subjects and trials. To this end, we compute the visual saliency for all the trials and subjects specific to each intention by selecting five equispaced frames from each trial and averaged all the corresponding saliency maps. Taking a closer look at the saliency maps overlapped to a sample video frames in Fig. 8, it is worth noting that, since in the experimental setup all the context is uninformative, the approach of Bargal et al. (2018) is focusing on the unique parts that are moving (hand, arm), and it is hard to extract general patterns specific to a single intention. However, at a finer level, some observations can be made. For example, it can be noted that the gestures finalized to Drinking and Placing appear to be more discriminative in the elbow movement on average. Instead, for Pouring and Passing the saliency maps seem mainly focusing on the wrist movement. Grounding on all these considerations, we have another confirmation of the difficulty of the task of predicting intention from motion, and how the grasping acts are actually visually similar even if the underlying intention is different. As compared to subject-specific saliency maps (Fig. 7), the integral map (Fig. 8) appears a bit more spread, suggesting that there exist subtle differences in how different subjects execute the grasping, even when finalized to the same intention. The presence of such extremely subtle differences can explain why SADA (and blind-SADA) is capable of boosting intention prediction by adapting across subjects.

6.2 Mining Discriminants for Intentions at the Feature Level

To mine intention-related discriminant in the grasping motor acts, we also pursued an analysis at the feature-level from 3D data.

To this aim, we run a statistical analysis out of the hidden representations learnt with SADA on top of ker-COV features. More precisely, we took the 200-dimensional vectorial representations obtained using our proposed method and we fixed one particular intention at a time. Then, we computed the correlation ρ (across all trials in the dataset) between each component of the feature representations and a binary vector, indexed over the trials, whose entries are 1 if and only if that given trial belong to the fixed specific intention (and

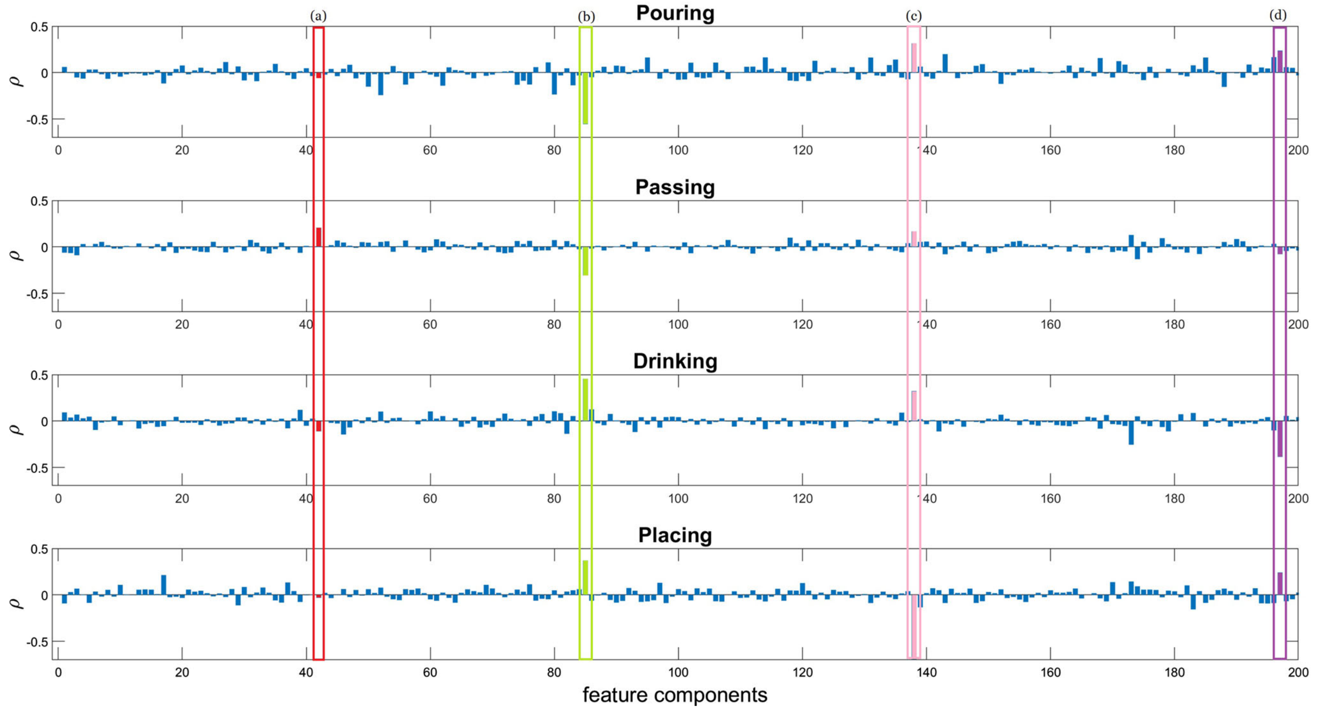


Fig. 9 Mining discriminants for intentions in high-level feature encodings from 3D data. We computed correlations values ρ for feature components and corresponding labels for intentions

we repeated the same procedure for all intentions). In this way, we can produce a global statistics over the IfM dataset and measure how much each component of the SADA representations is actually discriminative in favor of which single intention. Considering the well known statistical properties of correlation, we remind that values close to 1 or -1 indicate a strong positive or negative linear dependency for the given component in favor of a single intention.

The results of this approach are reported in Fig. 9. As one can see, there are some sharp peaks in the values of ρ and we highlighted some of them using specific colors. The component highlighted in red (a) seems to be mildly correlated with Passing and Drinking, with opposite signs. Similarly, the component in purple (d) shows a mild positive correlation with Pouring/Placing and a strong negative correlation with both Drinking and Passing. The component in pink (c) has a positive correlation with Drinking, Pouring and Passing, exhibiting a remarkably strong negative correlation with Placing (actually, we registered here the maximal negative correlation among all the feature components). Finally, the component in green (b) results in a strong dependency with all the intentions, showing a high positive correlations with Drinking/Placing and strong negative correlations with Pouring/Passing.

This suggests the fact that there is a relevant statistical variability of some feature components which are well aligned with the actual intentions' label to be predicted. Therefore, within the learned classification model, there are components which are more effective than others in encoding the kinematics of the grasping in order to allow the prediction of the underlying intention. Despite this feature-based analysis is less interpretable with respect the previous one based on frames, it suggests that intentions' discriminants are actually embedded and can be extracted from the whole kinematics of the onset movement.

6.3 Mining Discriminants for Intentions in 3D Data

In this Section we aim at analyzing interpretable kinematic descriptors, which precisely encode the 3D raw joints coordinates, for the sake of mining discriminants for intentions.

We considered the raw 3D data acquired from VICON system and we performed the average over a window whose length is 10% of the actual graspings' duration, quantizing it into 10 disjoint time segments, that is, 0-10%, 10%-20%, and so on. For each time segment, we computed low-level kinematic descriptors, to either encode absolute and relative geometrical configurations of the hand, as well as kinematic cues. Precisely, among relative geometrical descriptors, we computed the orientation of the palm (estimated through the direction of its normal), the grip aperture (distance between the markers on top of index and thumb), and the hand aperture (distance between the markers on top of index and baby finger). Absolute geometrical configuration were estimated by accounting the (x, y, z) coordinate of the centroid of each finger. Finally, we computed hand velocity and acceleration as kinematic descriptors. Such descriptors are well rooted in the the psychological literature Cavallo et al. (2016) for the sake of a kinematic analysis. Choosing such descriptors, we aim at providing "human interpretable" indicators to explain our computational findings, being well aware that these are not theoretically principled proofs, but mainly data-driven evidence that intentions can be predicted by the anticipating action. Afterwards, for each time window and descriptor, we computed mean and standard deviation across all trials in the IfM 3D dataset which belong to a fixed intention. Hence, we can approximate the geometrical area of influence of each intention with a 3D ellipsoid, using the mean as center and standard deviation as length for the semi-axes. The overlap between intentions is evaluated through the intersection over union (IoU) of the volumes of the ellipsoids. The IoU metric spans the range [0,1] in which the extremal cases 0 and 1 correspond to the maximal disambiguation and confusion between intentions, respectively. As evidence of a reliable prediction of intentions from motion, we expect to register low values from such metric. In fact, this will correspond to a favorable geometrical configuration in which intentions are very different from each others (up to a statistical approximation of the second order). The aforementioned geometrical configuration will translate into more reliable discriminants for intentions, which can be exploited in a subsequent classification stage, not considered here.

The results of our analysis are summarized in Fig. 10. When either accounting for relative, absolute geometry or kinematic low-level descriptors, we have often estimated an IoU metric that is often below the 0.5 threshold, being therefore statistically relevant. Even when considering the very initial time segment in which we quantized the graspings' duration, we could find statistical evidences for the presence of kinematic patterns which anticipate intentions. At the same time, as long as the grasping execution approaches its end, descriptors are less overlapped, denoting in principle an easier task for disambiguating intentions. In particular, when both considering absolute geometrical and kinematic descriptors, we can appreciate a sharp reduction of the overlap between intentions in the second half of the duration. Moreover, when comparing all three classes of low-level descriptors adopted, we have found that relative geometry descriptors (blue plot, left) provide a slightly worse separation results if compared to absolute geometry (green plot, center) and kinematic (red plot, right) descriptors, mainly during the early stage of the grasping. On the one hand, this can be explained as the effect of the initial fixed resting position of the hand, which is constant across intentions. However, such starting position is almost immediately varied in favor of a hand's configuration which is specific for the overarching intent: this explains why later time segments show a reduced overlap between intentions. On the other hand, relative descriptors are independent from the frame of reference and thus inherently invariant to roto-translations of the arm, showing thus less variability across intentions (i.e., more statistical overlap, especially in the early stages of motion). Another consideration comes from the analysis of finger positions. Among the descriptors which codify the absolute geometry of the hand, the thumb and the index are slightly better than the other fingers in reducing the overlap between fingers, registering an IoU less than 0.1 in the last temporal segment. Therefore, we can assert that there is evi-

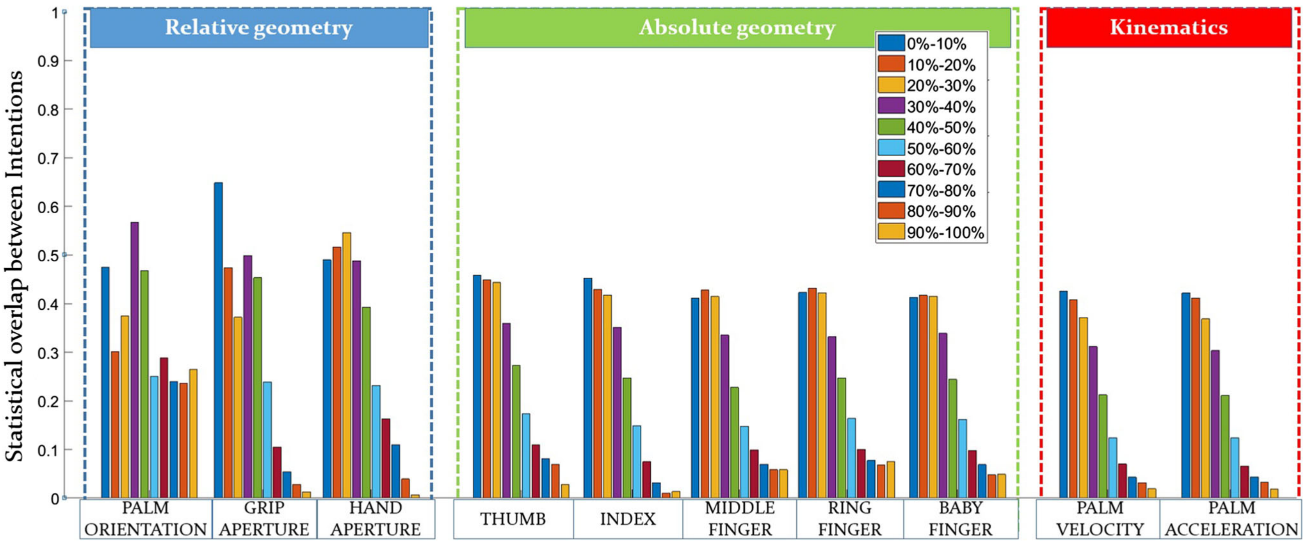


Fig. 10 We reported intersection over union (IoU) values to compare the degree of separability of intention-specific ellipsoids, built by accounting first and second order statistics of the raw 3D joints coordinates, as well as interpretable kinematic features inspired by Cavallo et al. (2016). Lower IoU values correspond to a better separability between intentions

dence that the position of the hand is prepared in advance and accommodated in a way that, according to our analysis, is specific for each intention. A similar level of performance is achieved by hand velocity and acceleration, furthermore proving that intentions can be predicted from the kinematics only.

6.4 Discussion

If we combine the findings obtained from the different types of analyses reported in Sects. 6.1, 6.2 and 6.3, we can observe complementary aspects. At the level of video data (Figs. 7 and 8), when attempting to locate in space and time the actual discriminants for the intention to be predicted, there does not seem to emerge any evident visual explanation. This is also reasonable if we consider that predicting intentions from motion only is an arguably challenging task, which becomes even harder when considering the additional request of finding a specific spatio-temporal segment that explains the prediction in a standalone manner. However, our classification performance shows that anticipating human intentions is actually feasible and, therefore, such discriminants do exist, but reliably extracting them is not trivial. This is a problem that can be tackled by looking at feature encodings (Fig. 9), i.e., considering the intention-specific patterns discovered inside the feature representations obtained through subject-adversarial training. In fact, we can argue that intention-related discriminants are available inside the grasping motor act, and they become evident when the complete kinematics is encoded. We showed that by using simple features to encode relative/absolute positions and dynamical states of the hand during the execution of the grasping, it is possible to appreciate that later instants convey more geometric separability among intentions (Fig. 10), but the whole kinematics does matter, even from the very onset of the movement. Ultimately, this enriches the apparently unrelated motor act with patterns which can be exploited to predict intentions by computational methods, being nevertheless fairly difficult to be visualized.

7 Conclusions and Future Work

In this paper, we introduce Intention from Motion, a novel problem consisting in predicting the goal i.e., intention) that originates from an human action by using the kinematics only, in a context-free setting. We have presented a new dataset and found that by only inspecting grasping-a-bottle actions, we can predict whether they fulfill a Pouring, Passing, Drinking or Placing intention.

As the result of a broad baseline analysis, we proved that our novel problem is feasible and intention discriminants are embedded in the anticipative and apparently unrelated grasping motor act. We also demonstrated that, as to ensure those discriminants to be generalizable across subjects, a domain adaptation technique is proposed and proficiently applied to our intention prediction scenario and to standard action recognition settings as well. When interpreting each subject as a domain, Subject-Adversarial Domain Adaptation (SADA) remarkably boosts the prediction capability for intentions.

As an extension, we proposed Blind-SADA to show that exploiting subjects' identities only in training to perform adaptation leads to good generalization on an unknown agent. Despite less data are exploited by Blind-SADA, its performance is not too far degraded from the one of SADA, and both improve upon the baseline. This certifies the effectiveness of our idea of learning from multiple subjects as to adapt on both specific and general target domains/subjects. Finally, we have provided some insights on which intentions differentiators can be extracted from video sequences and/or 3D joints coordinates.

Future directions will consider the extension of this frameworks towards applications in robotics (e.g., human–robot interaction) and video-surveillance in which more complicated actions, different objects and composite contexts will be considered. Also, it will be interesting to scale the prediction of intentions in a full-body setup, investigating the usage of more portable devices for skeletal joints acquisition, such as depth sensors.

References

Ansuini, C., Cavallo, A., Bertone, C., & Becchio, C. (2014). Intentions in the brain: The unveiling of mister hyde. *The Neuroscientist*, *21*, 126–135.

Ansuini, C., Cavallo, A., Koul, A., Jacono, M., Yang, Y., & Becchio, C. (2015). Predicting object size from hand kinematics: a temporal perspective. *PloS ONE*, *2*(10), e0120432.

Bargal SA, Zunino A, Kim D, Zhang J, Murino V, Sclaroff S (2018) Excitation backprop for RNNs. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Becchio, C., Koul, A., Ansuini, C., Bertone, C., & Cavallo, A. (2018). Seeing mental states: An experimental strategy for measuring the observability of other minds. *Physics of Life Reviews*, *24*, 67–80.

Bütepage J, Black MJ, Kragic D, Kjellström H (2017) Deep representation learning for human motion prediction and classification. In *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)*.

Cao Y, Barrett D, Barbu A, Siddharth N, Yu H, Michaux A, Lin Y, Dickinson S, Siskind JM, Wang S (2013) Recognizing human activities from partially observed videos. In *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)*.

Cavallo A, Koul A, Ansuini C, Capozzi F, Becchio C (2016) Decoding intentions from movement kinematics. Scientific Reports

Cavazza J, Zunino A, San Biagio M, Murino V (2016) Kernelized covariance for action recognition. In *Proceedings of the IEEE international conference on pattern recognition (ICPR)*.

Cavazza J, Morerio P, Murino V (2017a) A compact kernel approximation for 3d action recognition. In International Conference on Image Analysis and Processing (ICIAP)

Cavazza J, Morerio P, Murino V (2017b) When kernel methods meet feature learning: Log-covariance network for action recognition from skeletal data. In *Proceedings IEEE Conference on computer vision and pattern recognition workshop (CVPRw)*.

Cavazza, J., Morerio, P., & Murino, V. (2019). Scalable and compact 3D action recognition with approximated RBF kernel machines. *Pattern Recognition*,. https://doi.org/10.1016/j.patcog.2019.03.031.

Chakraborty A, Roy-Chowdhury K (2014) Context-aware activity forecasting. In *Proceedings Asian conference on computer vision (ACCV)*.

Chopra S, Balakrishnan S, Gopalan R (2013) Dlid: Deep learning for domain adaptation by interpolating between domains. In *ICML workshop on challenges in representation learning*

Csurka G (2017) A comprehensive survey on domain adaptation for visual applications. In *Domain adaptation in computer vision applications* (pp. 1–35). Springer,

Daumé, H, I. I. I., & Marcu, D. (2006). Domain adaptation for statistical classifiers. *Journal of Artificial Intelligence Research*, *26*(1), 101–126.

Davis, J. W., & Tyagi, A. (2006). Minimal-latency human action recognition using reliable-inference. *Image and Vision Computing*, *24*(5), 455–472.

Fermüller, C., Wang, F., Yang, Y., Zampogiannis, K., Zhang, Y., Barranco, F., et al. (2017). Prediction of manipulation actions. *International Journal of Computer Vision*, *126*, 358.

Fernando B, Habrard A, Sebban M, Tuytelaars T (2013) Unsupervised visual domain adaptation using subspace alignment. In *Proceedings IEEE international conference on computer vision (ICCV)*.

Ganin Y, Lempitsky V (2015) Unsupervised domain adaptation by backpropagation. In *Proceedings international conference on machine learning (ICML)*.

Ganin, Y., Ustinova, E., Ajakan, H., Germain, P., Larochelle, H., Laviolette, F., et al. (2016). Domain-adversarial training of neural networks. *Journal of Machine Learning Research*, *17*(59), 1–35.

Ghifary M, Kleijn WB, Zhang M (2014) Domain adaptive neural networks for object recognition. CoRR abs/1409.6041

Gong B, Shi Y, Sha F, Grauman K (2012) Geodesic flow kernel for unsupervised domain adaptation. In *Proceedings IEEE conference on computer vision and pattern recognition (CVPR)*.

Gopalan R, Li R (2011) Domain adaptation for object recognition: An unsupervised approach. In *Proceedings IEEE international conference on computer vision (ICCV)*.

Hoai M, De la Torre F (2012) Max-margin early event detectors. In *Proceedings IEEE conference on computer vision and pattern recognition (CVPR)*.

Huang D, Wang YF (2013) Coupled dictionary and feature space learning with applications to cross-domain image synthesis and recognition. In *Proceedings IEEE international conference on computer vision (ICCV)*.

Huang DA, Kitani KM (2014) Action-reaction: Forecasting the dynamics of human interaction. In *Proceedings European conference on computer vision (ECCV)*.

Jain A, Zamir AR, Savarese S, Saxena A (2016) Structural-RNN: Deep learning on spatio-temporal graphs. In *Proceedings IEEE conference on computer vision and pattern recognition (CVPR)*.

Jégou H, Douze M, Schmid C, Pérez P (2010) Aggregating local descriptors into a compact image representation. In *Proceedings IEEE conference on computer vision and pattern recognition (CVPR)*.

Kilner, J. M. (2011). More than one pathway to action understanding. *Trends in Cognitive Sciences*, *15*, 352–357.

Kitani KM, Ziebart BD, Bagnell JA, Hebert M (2012) Activity forecasting. In *Proceedings European conference on computer vision (ECCV)*.

Kong, Y., & Fu, Y. (2016). Max-Margin Action Prediction Machine. *IEEE Trans on Pattern Analysis and Machine Intelligence (PAMI)*, *38*(9), 1844–1858.

Kong Y, Tao Z, Fu Y (2017) Deep sequential context networks for action prediction. In *Proceedings IEEE conference on computer vision and pattern recognition (CVPR)*.

Koppula, H. S., & Saxena, A. (2016). Anticipating human activities using object affordances for reactive robotic response. *IEEE Transaction on Pattern Analysis and Machine Intelligence (PAMI)*, *38*(1), 14–29.

Koul, A., Cavallo, A., Cauda, F., Costa, T., Diano, M., Pontil, M., et al. (2018). Action observation areas represent intentions from subtle kinematic features. *Cerebral Cortex*, *28*(7), 2647–2654.

Koul, A., Soriano, M., Tversky, B., Becchio, C., & Cavallo, A. (2019). The kinematics that you do not expect: Integrating prior information and kinematics to understand intentions. *Cognition*, *182*, 213–219.

Kuehne H, Jhuang H, Garrote E, Poggio T, Serre T (2011) Hmdb: A large video database for human motion recognition. In *Proceedings IEEE international conference on computer vision (ICCV)*.

Lan T, Chen TC, Savarese S (2014) A hierarchical representation for future action prediction. In *ECCV*

Li, K., & Fu, Y. (2014). Prediction of human activity by discovering temporal sequence patterns. *IEEE Transaction on Pattern Analysis and Machine Intelligence (PAMI)*, *36*(8), 1644–1657.

Li K, Hu J, Fu Y (2012) Modeling complex temporal composition of actionlets for activity prediction. In *Proceedings European conference on computer vision (ECCV)*.

Li W, Zhang Z, Liu Z (2010) Action recognition based on a bag of 3d points. In *Proceedings IEEE conference on computer vision and pattern recognition workshop (CVPRw)*.

Liu MY, Tuzel O (2016) Coupled generative adversarial networks. In *Proceedings advances in neural information processing systems (NIPS)*.

Lu C, Hirsch M, Schölkopf B (2017) Flexible spatio-temporal networks for video prediction. In *Proceedings IEEE conference on computer vision and pattern recognition (CVPR)*.

Ma S, Sigal L, Sclaroff S (2016) Learning activity progression in LSTMs for activity detection and early detection. In *Proceedings IEEE conference on computer vision and pattern recognition (CVPR)*.

Moeslund, T. B., Hilton, A., & Krüger, V. (2006). A survey of advances in vision-based human motion capture and analysis. *Computer Vision and Image Understanding (CVIU)*, *104*(2), 90–126.

Morerio P, Cavazza J, Murino V (2018) Minimal-entropy correlation alignment for unsupervised deep domain adaptation. In *Proceedings international conference on learning representations (ICLR)*.

Müller M, Röder T, Clausen M, Eberhardt B, Krüger B, Weber A (2007) Documentation mocap database HDM-05. Tech. Rep. CG-2007-2, Universität Bonn

Oztop, E., Wolpert, D., & Kawato, M. (2005). Mental state inference using visual control parameter. *Cognitive Brain Research*, *22*, 129–151.

Rohrbach M, Amin S, Andriluka M, Schiele B (2012) A database for fine grained activity detection of cooking activities. In *Proceedings IEEE conference on computer vision and pattern recognition (CVPR)*.

Ryoo MS (2011) Human activity prediction: Early recognition of ongoing activities from streaming videos. In *Proceedings IEEE international conference on computer vision (ICCV)*.

Ryoo MS, Fuchs TJ, Xia L, Aggarwal JK, Matthies L (2015) Robot-centric activity prediction from first-person videos: What will they do to me? In *Proceedings IEEE international conference on human–robot interaction (HRI)*.

Schindler K, Gool LJV (2008) Action snippets: How many frames does human action recognition require? In *Proceedings IEEE conference on computer vision and pattern recognition (CVPR)*.

Shekhar S, Patel VM, Nguyen HV, Chellappa R (2013) Generalized domain-adaptive dictionaries. In *Proceedings IEEE conference on computer vision and pattern recognition (CVPR)*.

Soomro K, Zamir AR, Shah M (2012) Ucf101: A dataset of 101 human action classes from videos in the wild. In *CRCV-TR-12-01*.

Soomro K, Idrees H, Mubarak S (2016) Predicting the where and what of actors and actions through online action localization. In *Proceedings IEEE conference on computer vision and pattern recognition (CVPR)*.

Soran B, Farhadi A, Shapiro L (2015) Generating notifications for missing actions: Don’t forget to turn the lights off! In *Proceedings IEEE international conference on computer vision (ICCV)*.

Soriano, M., Cavallo, A., D?Ausilio, A., Becchio, C., & Fadiga, L. (2018). Movement kinematics drive chain selection toward intention detection. *Proceedings of the National Academy of Sciences*, *115*(41), 10452–10457.

Stapel JC, Hunnius S, Bekkering H (2012) Online prediction of others’ actions: the contribution of target object, action context, and movement kinematics. In Psychological Research

Stein S, McKenna SJ (2013) Combining embedded accelerometers with computer vision for recognizing food preparation activities. In *Proceedings ACM international joint conference on pervasive and ubiquitous computing*.

Sun B, Saenko K (2016) Deep CORAL: correlation alignment for deep domain adaptation. In *Proceedings European conference on computer vision (ECCV)*.

Sun B, Feng J, Saenko K (2016) Return of frustratingly easy domain adaptation. In *Proceedings AAAI conference on Artificial Intelligence (AAAI)*.

Taigman Y, Polyak A, Wolf L (2017) Unsupervised cross-domain image generation. In *Proceedings international conference on learning representations (ICLR)*.

Tzeng E, Hoffman J, Zhang N, Saenko K, Darrell T (2014) Deep domain confusion: Maximizing for domain invariance. CoRR abs/1412.3474

Tzeng E, Hoffman J, Darrell T, Saenko K (2015) Simultaneous deep transfer across domains and tasks. In *Proceedings IEEE international conference on computer vision (ICCV)*.

Tzeng E, Hoffman J, Darrell T, Saenko K (2017) Adversarial discriminative domain adaptation. In *Proceedings IEEE conference on computer vision and pattern recognition (CVPR)*.

Van der Maaten, L., & Hinton, G. (2008). Visualizing data using t-SNE. *Journal of Machine Learning Research (JMLR)*, *9*(2579–2605), 85.

Volpi R, Morerio P, Savarese S, Murino V (2018) Adversarial feature augmentation for unsupervised domain adaptation. In *Proceedings IEEE conference on computer vision and pattern recognition (CVPR)*.

Vondrick C, Pirsiavash H, Torralba A (2016) Anticipating visual representations with unlabeled video. In *Proceedings IEEE conference on computer vision and pattern recognition (CVPR)*.

Walker J, Gupta A, Hebert M (2014) Patch to the future: Unsupervised visual prediction. In *Proceedings IEEE conference on computer vision and pattern recognition (CVPR)*.

Wang, H., Kläser, A., Schmid, C., & Liu, C. L. (2013). Dense trajectories and motion boundary descriptors for action recognition. *International Journal of Computer Vision (IJCV)*, *103*(1), 60–79.

Xie C, Li C, Zhang B, Chen C, Han J, Liu J (2018) Memory attention networks for skeleton-based action recognition. In *Proceedings international joint conference on artificial intelligence (IJCAI)*.

Xu Z, Qing L, Miao J (2015) Activity auto-completion: Predicting human activities from partial videos. In *Proceedings IEEE international conference on computer vision (ICCV)*.

Zunino A, Cavazza J, Koul A, Cavallo A, Becchio C, Murino V (2017a) Predicting human intentions from motion cues only: A 2d+3d fusion approach. In *Proceedings ACM conference on multimedia*

Zunino A, Cavazza J, Koul A, Cavallo A, Becchio C, Murino V (2017b) What will i do next? The intention from motion experiment. In *Proceedings IEEE conference on computer vision and pattern recognition workshop (CVPRw)*.

Zunino A, Cavazza J, Murino V (2017c) Revisiting human action recognition: Personalization vs. Generalization. In *Proceedings international conference on image analysis and processing (ICIAP)*.

Zunino A, Morerio P, Cavallo A, Ansuini C, Podda J, Battaglia F, Veneselli E, Becchio C, Murino V (2018) Video gesture analysis for autism spectrum disorder detection. In *Proceedings IEEE international conference on pattern recognition (ICPR)*.

Publisher’s Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.