What Will I Do Next?
The Intention from Motion Experiment

Andrea Zunino[1,3], Jacopo Cavazza[1,3], Atesh Koul[2], Andrea Cavallo[2,4],
Cristina Becchio[2,4] and Vittorio Murino[1,5]

firstname.lastname@iit.it

[1] Pattern Analysis & Computer Vision (PAVIS) – Istituto Italiano di Tecnologia – *Genova, Italy*
[2] Robotics, Brain and Cognitive Science – Istituto Italiano di Tecnologia (IIT) – *Genova, Italy*
[3] Electrical, Electronics and Telecommunication Engineering and Naval Architecture Department (DITEN) – Università degli Studi di Genova – *Genova, Italy*
[4] Psychology Department – University of Torino – *Torino, Italy*
[5]Computer Science Department – Università di Verona – *Verona, Italy*

Abstract

In computer vision, video-based approaches have been widely explored for the early classification and the prediction of actions or activities. However, it remains unclear whether this modality (as compared to 3D kinematics) can still be reliable for the prediction of human intentions, defined as the overarching goal embedded in an action sequence. Since the same action can be performed with different intentions, this problem is more challenging but yet affordable as proved by quantitative cognitive studies which exploit the 3D kinematics acquired through motion capture systems.

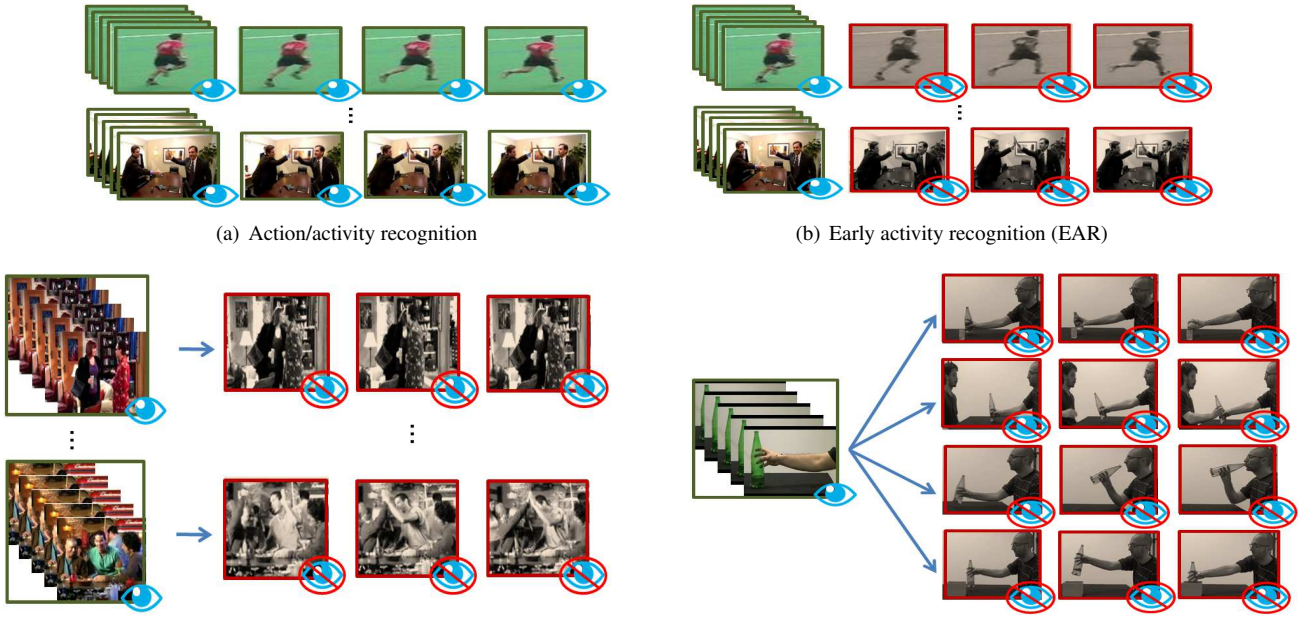
In this paper, we bridge cognitive and computer vision studies, by demonstrating the effectiveness of video-based approaches for the prediction of human intentions. Precisely, we propose Intention from Motion*, a new paradigm where, without using any contextual information, we consider instantaneous grasping motor acts involving a bottle in order to forecast why the bottle itself has been reached (to pass it or to place in a box, or to pour or to drink the liquid inside).*

We process only the grasping onsets casting intention prediction as a classification framework. Leveraging on our multimodal acquisition (3D motion capture data and 2D optical videos), we compare the most commonly used 3D descriptors from cognitive studies with state-of-the-art video-based techniques. Since the two analyses achieve an equivalent performance, we demonstrate that computer vision tools are effective in capturing the kinematics and facing the cognitive problem of human intention prediction.

1. Introduction

Action and activity recognition are one of the most active areas in computer vision. The task here consists in the classification of *fully observed* sequence and many methods have been proposed to tackle this task [17, 27]. More recently, the community has also started to investigate a few variants, either performing the early classification of partially disclosed activities or predicting future actions by analyzing the events occurring up to a certain instant. For the sake of clarity, let us briefly review these two different paradigms which are also sketched in Fig. 1.

Early Activity Recognition (EAR). Ryoo [20] devised a system to infer the ongoing activity by only analysing its *onset*, *i.e*. the initial part of the action. This is done with a dynamic programming method to match an extension of classical bag-of-features representation which allow to capture the temporal correlation of descriptors. Hoai and De la Torre [12] designed MMED, Max-Margin Early-event Detectors to address the problem to understand a specific human emotion after it starts but before it ends. They trained a structured output support vector machine [23] on the whole accomplished emotion development and, during testing phase, they managed to classify fear rather than disgust when they are about to occur. Yu et al. [30] propose a generalization of Spatio-Temporal Interest Point [17] to categorize actions from their beginning. The inter-dependencies between different spatial location are implemented into a probabilistic graphical model fed by histogram features. Cao et al. [6] split a complete action into temporal segments which are further represented by means of sparse coding, so that actions are recognizable from incomplete data. Ryoo



(a) Action/activity recognition

(b) Early activity recognition (EAR)

(c) Action prediction (AP)

(d) **Intention prediction**

Figure 1. Four different paradigms in human action/activity analysis. 1(a) Action/activity recognition: each sequence is fully observed to infer the class label ("running" for the top sequence up to "high-five" for the bottom). 1(b) Early activity recognition: only a few initial frames per sequence is observed and the goal is an early classification from these incomplete observations. 1(c) Action prediction: future actions are predicted analysing all past events which, in general, can be very different across different classes. Thus, in the top sequence a standing up activity leads to predict a "kissing", while, in the bottom, a conversation between a group of friends anticipates a "high-five". 1(d) Intention prediction: a novel paradigm where unobserved future action are anticipated from the same class of motor act, all extremely similar in appearance, no matter which different ending will occur.

et al. [21] tackle early activity recognition from egocentric videos: the task is detecting the so-called *onset signature*, a bunch of kinematic evidence which has strong predictive properties about the last part of the observed action. Some works have attempted to investigate how much of the whole action is necessary to perform a classification: Davis and Tyagi [10] adopt a generative probabilistic framework to deal with the uncertainty due to limited amount of data, while Schindler and Van Gool [22] try to answer the aforementioned question using a similarity measure between the statical and the motion information extracted from videos.

Action Prediction (AP). Lan and Savarese [16] developed the so-called *hierarchical movemes*, a new representation to model human actions from coarse to fine levels of granularities which was integrated in a max-margin learning framework for action prediction. Vondrick and Torralba [25] uses a deep neural network trained over 600 hours of videos. During training the net exploits videos to learn to predict the representation of frames in the future and the last fully connected layer allows to perform classification over different future endings. For the sake of prediction, contextual information plays a pivotal role. Indeed, once the objects present in a scene are detected, the object-object or object-person relationship can be modelled by several probabilistic architectures (*e.g.*, graphical models [9, 11] or topic models [15, 19]). Among the works which directly models the context inside the algorithms, some of them deal with the prediction of future trajectories of moving objects (vehicles or pedestrian) [14, 26, 29, 28] by estimating the spatial areas over which such objects will most likely pass with respect to those which are excluded by this passage (*e.g.*, car circulations over sidewalks [26]).

Shortcomings in EAR and AP. Previous attempts in EAR and AP frequently exploit motion patterns which are specific of the subsequent actions, since they contain some cues that undoubtedly help the recognition. For instance, if the goal is understanding whether two people are going to shake their hands or to give a high-five, by just looking at the first part of their interaction, a low wrist height can be an evidence of a handshaking [25, 16]. Further, another important aspect of the entire activity recognition problem is that the current techniques are mainly exploiting the scene context to support the classification ([6, 11, 9, 28, 26, 19, 9, 13, 24], and [5]), i.e., the objects present in the scene and the knowledge about the actions associated to them are cues that can be utilized to help in

making a correct inference of the ongoing action to be recognized. This information is necessary but often insufficient to solve the issue or, worse, the context may not always be available or easily recognizable, being also misleading when the scene is too noisy or cluttered [31]. In any case, an important source of information to disambiguate intention can be provided by the kinematics of the movement.

Cognitive studies. Recent findings in cognition indicate that how a motor act is performed (e.g., grasping an object) is not solely determined by biomechanical constraints imposed by the object extrinsic and intrinsic properties with which one is interacting, but depends on the agents intention. Indeed, intentions become "visible" in the displayed actions of agents manipulating a given object [2, 4], and this process is actually readable by observers. Manera et al. [18] showed that in a binary choice design, observers were able to judge whether the agents intent in grasping the object was to cooperate with a partner or compete against an opponent. In addition, by means of a motion capture system to acquire the exact kinematics, Ansuini et al. [2] proved that how a given object is grasped is richly informative for observer to determine the social vs. individual agent's intention. Recently, Becchio et al. [8] demonstrated that it is possible to quantify the impact of kinematic variables (such as wrist height/velocity) for the sake of intentions' recognition.
The analysis is extremely interesting since may corroborate the actual feasibility of predicting intentions even in context-free settings, where kinematics is the only available source of information. Nevertheless, there is a gap between quantitative evidences drawn from 3D kinematics [2, 8] and human performance analysis carried out on video data [18]. Specifically, it is still unclear whether the subtle kinematics which differentiate intentions is also exploitable from RGB videos in a quantitative fashion.

1.1. Paper Contributions

In this paper, we aim at introducing **Intention from Motion** (IfM), a brand new problem with two challenging aspects largely differentiate this work from the current literature of either EAR or AP.

- Grounding from the assumption that the same class of motor acts can be performed with different intentions, we want to analyse the movement onset of an apparently unrelated action (actually embedding the intention from the very beginning), the same for all intentions, capturing those subtle motion patterns which are anticipative of the future action.
- Unlike the main existing literature, we want to avoid the exploitation of any hints derived by the context, solely focusing on the kinematics of the movement.

This novel setup was accomplished by a set of experiments where subjects were asked to grasp a bottle, 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 box. The dataset is composed by both 3D trajectories of twenty motion capture (VICON) markers outfitted over the hand of the participants and optical RGB video sequences lasting about one second, with an occlusive camera view in which only the arm and the bottle are visible. Data are acquired from the moment when the hand starts from a stable fixed position up to the reaching of the object, and 3D marker trajectories and video sequences are exactly trimmed at the instant when the hand grasps the bottle, removing the following part. The goal is to classify the intentions associated with the observed grasping-a-bottle movement, *i.e*. to predict the agent's intention.

Even if some methodologies have been proposed to for EAR and AP [16, 20, 25, 12, 26], the experiments are typically performed on standard action recognition datasets, just adapted to the new task, and often considering the start of the same action which of course helps. On the contrary, our experiment is explicitly designed for intention prediction.

Due to the multimodality of our experimental settings, our work interleaves the EAR and AP problems from computer vision with the kinematic approach of cognitive studies in predicting intentions. Indeed, by taking advantage of state-of-the-art hand-crafted features in video-based action recognition, we are able to prove that the subtle kinematic analysis required to predict intentions is actually feasible while leveraging on RGB optical video as an alternative approach to motion capture data [2, 8] which are obviously more difficult to obtain.

In summary, the present work introduces the following main contributions.

(a) We introduce the new problem of Intention from Motion. That is, from the same observable "neutral" motor act - used in both training and test phases - we try to classify the underlying intention using solely motion information, without any contextual cue. Unlike previous works, we are neither classifying actions from their very first beginning (e.g., [20, 12]) nor classifying different futures by analyzing different past onsets [16, 25]. Instead, we anticipate intentions which finalize the same class of motor act, distilling from it the discriminative motion patterns characterizing the specific intention while fully neglecting any contextual information (as opposed to [9] or [26]).

(b) We propose a (3D + 2D) dataset specifically aimed at the prediction of human intentions. This dataset is designed in a principled way by defining four intentions (Pouring, Passing, Drinking, Placing) performed by independent naive subjects, which are all forerun from the very similar initial movement of grasping-a-bottle, while avoiding bias which can affect the subsequent

performance analysis. To the best of our knowledge, this is the first time a dataset has been explicitly designed for intention prediction.

(c) We carry out a separate 3D and 2D analysis, exploiting either broadly used 3D kinematic features or state-of-the-art video-based approaches. In light of the equivalent performance obtained by the two, we bridge the gap between video-based approaches for early activity recognition/action prediction (computer vision) and the analysis of 3D kinematics acquired by motion capture (cognitive studies).

Paper outline. Section 2 introduces our dataset and the experimental setting. We investigate human performance in intention prediction in Section 3. Subsequently, we extensively describe the experimental analysis accomplished on the 3D data in Section 4, and on the 2D video sequences in Section 5. Section 6 finally draws the conclusions.

2. Dataset overview

Seventeen naive volunteers were seated beside a 110×100 cm table resting on it elbow, wrist and hand inside a 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 $17 \times 17 \times 12.5$ box positioned on the same table, 25 cm distant.

After a preliminary session, in which participants are familiarized with the execution, each subject performed 20 trials per intention. The experimenter visually monitored each trial to ensure exact compliance of these requirements. In order to homogenize the dataset, we completely removed trials judged imprecise. Thus, the final dataset includes 1098 trial (253 for pouring, 262 for passing, 300 for drinking and 283 for placing) and, for each of them, both 3D and video data have been collected.

3D kinematic data. Near-infrared 100 Hz VICON system was used 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. 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 represented with a set of

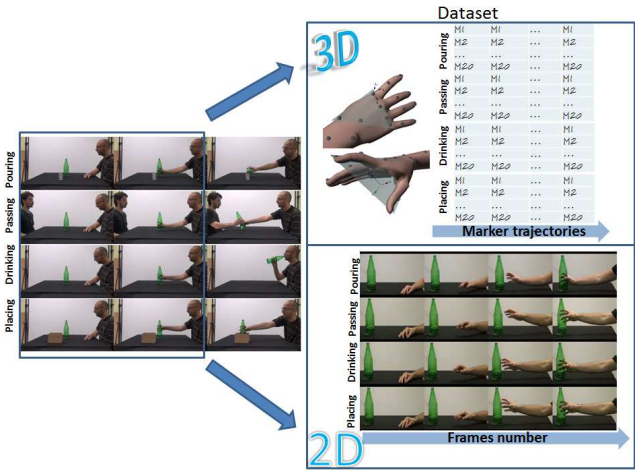


Figure 2. The dataset. On the left we have the entire visible pouring, passing, drinking and placing development. On the top right, we have 3D VICON data acquisition, on bottom right, video sequences in which camera shoots only the arm and the bottle. In both cases, the acquisition stop at the grasping moment.

3D points describing the trajectory covered by every single marker during execution phase. The x, y, z marker coordinates only consider the reach-to-grasp phase, where 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. Movements 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 filmed. 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 understanding 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. Consequently all the sequences result about 30 frames long (some sequences are available in the Supplementary Material).

In all the experiments reported in this paper, either dealing with 3D or 2D data, we consider all the possible pairwise comparisons between intentions and the all-class one.

We select one-subject-out testing procedure. That is, we compute seventeen accuracies, training our system on all the subjects except the one we are testing, then we averaged all the accuracies to get the final classification results.

3. Human performance in IfM

As a preliminary analysis to check how human beings can predict intentions, we tested the human capabilities on Pouring vs. Placing and Pouring vs. Drinking throughout the following experimental apparatus. We asked each of 36 participants to watch 400 videos of reach-to-grasp movements and predict if 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 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. 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 vs. Placing test, we get 68% of accuracy. Afterward, we move to the second test (Pouring vs. Drinking) and we repeated the same procedure. In Pouring vs. Drinking, accuracy decreases to 58% (-10%). These results are statistically significant and suggest that human observers are able to exploit kinematics to predict intentions. Globally, although the human brain can read in a grasping some motion pattern which anticipates four different intentions, the computer vision methods explained in the paper are more valuable and outperforming human predictive ability.

4. 3D kinematic analysis

Recent social cognitive work [3] argued that intentions become "visible" in the apparent motion flow and understanding other's intentions cannot be divorced from the discrimination of essential kinematics. Thus, we exploited 3D kinematic features (KF) for IfM. Following [7], we computed *wrist velocity*, the module of the velocity of the wrist marker, *wrist height*, the z-component of the wrist marker, *wrist horizontal trajectory* defined as the x-component of the wrist marker and *grip aperture*, *i.e*. the distance thumb-index tips markers. Such features were referred to the reference system of the motion capture system, F_{global} [7]. A better characterization of the dynamics can be provided using a local reference system centered on the hand, F_{local} [4]. In this way, we computed relative x, y, z coordinates of thumb, index, thumb-index plane and the radius-phalanx. These variables provide the information about either the adduction/abduction movement of the thumb and index fingers or the rotation of the hand dorsum. Thus, they ensure robustness towards finger flexion/extension or wrist rotation that can vary significantly from one trial to another [4]. The 4 features from F_{global} and the 12 from F_{local} gives a total amount of 16 kinematic features. Acquisition time $[t_0, t_f]$ (see Section 2) is scaled into $[0, 1]$ and data are sub-sampled with step 0.01. Consequently, for each of our 16 kinematic features, we have 100 equispaced values describing the evolution of such features during the reach-to-grasp movement.

	Linear SVM fed with KF		
	F_{local}	F_{global}	F_{k}
Pouring vs. Placing	79,70	86,10	84,32
Pouring vs. Drinking	72,15	70,36	76,48
Pouring vs. Passing	76,55	67,39	**82,81**
Passing vs. Drinking	63,10	68,05	70,75
Passing vs. Placing	62,60	64,38	69,44
Drinking vs. Placing	64,40	71,41	73,72
All-class	45,08	48,01	55,13

Table 1. Results from 3D data. For the kinematic features (KF), a linear $C = 10$ SVM is fed with F_{local} and F_{global} groups of features, as well as with their combination F_{k}.

In Table 1, we report the classification results, using a linear support vector machine (SVM), considering F_{local} and F_{global} [7] individually, then concatenating all the features in F_{k} [4]. In the three cases, we used a feature vector of 1200, 400 and 1600 components respectively. As expected, when we combine F_{local} and F_{global}, the performance generally improves. Our results show that KF are able to obtain classification performances with a substantial improvement over the random guess level: using F_{k}, an average +26,25% improvement over the random guess level on the pairwise comparisons and +20,13% on the all-class case. Relying on the kinematic interpretability of KF, we conclude that the actual dynamic of the grasping encodes some motion patterns which go beyond the fulfilment of the action itself and can concretely anticipate the underlying intention.

5. Video-based analysis

To investigate the affordability of IfM we take advantage of some of the most effective spatio-temporal techniques [17, 27]. We compare STIP [17] and dense trajectories [27], which will be shortened in Section 5.2. In Section 5.3 we perform an analysis considering only fragments of video frames. Finally, in Section 5.4, a discussion of the achieved results is reported.

5.1. Local spatio-temporal features for IfM

To perform action recognition, the class of approaches, named in [1] as local, extracts some interest points (IPs),

Comparisons	$\mathcal{S}$=600	$\mathcal{S}$=1000	$\mathcal{S}$=2000
Pouring vs. Placing	79,69	**81,22**	79,41
Pouring vs. Drinking	**64,62**	64,15	64,58
Pouring vs. Passing	59,73	62,48	**62,63**
Passing vs Drinking	**59,53**	58,43	57,00
Passing vs Placing	56,87	59,96	**60,22**
Drinking vs Placing	65,97	**67,84**	66,61
All 4 intentions	36,64	**39,07**	38,41

Table 2. STIP classification percentages.

Comparisons	$\mathcal{S}$=600	$\mathcal{S}$=1000	$\mathcal{S}$=2000	$\mathcal{S}$=5000	$\mathcal{S}$=10000
Pouring vs. Placing	82,65	82,76	**83,18**	82,45	82,99
Pouring vs. Drinking	69,97	70,50	**71,73**	70,79	71,10
Pouring vs. Passing	74,85	74,79	75,61	75,55	**75,87**
Passing vs Drinking	67,20	66,62	68,37	**68,95**	67,98
Passing vs Placing	68,00	**68,46**	67,16	68,00	68,25
Drinking vs Placing	67,58	68,39	70,39	70,12	**70,50**
All 4 intentions	46,08	46,76	47,24	47,18	**47,35**

Table 3. Dense trajectories classification percentages $\mathcal{S} = 600, \ldots, 10000$.

detecting remarkable variations in both space and time, and associate each of them to a volume from which features are computed. Among the most effective methods, STIP [17] and dense trajectories (DT) [27] use a *sparse* or *dense* approach, respectively, to extract IPs. For STIP [17], a convolution with a Gaussian filter is used, while for DT [27], at different scales, a dense grid of points is initialized and each of them is tracked using optical flow. Subsequently, spatio-temporal volumes are generated by stacking the $N \times N$ spatial neighbours centered in all the L points in any tracked trajectory. The volume is warped in order to follow the related trajectory and it is subdivided in a $n_x \times n_y \times n_t$ grid of cuboids. For any channel of features, a dictionary of visual words is created: if B denotes the vocabulary size, each video v is represented by a set of histograms H^i, where i runs over the channels. We have $H^i = \{h^i_1, \ldots, h^i_B\}$, where $h^i_b > 0$ for any $b = 1, \ldots, B$ and we normalized $\sum_b h^i_b = 1$ for any i. For the classification, the following exponential χ^2 kernel is adopted.

$$K(v, \overline{v}) = \exp\left(-\frac{1}{2}\sum_i \frac{d(H^i, \overline{H}^i)}{A_i}\right), \qquad (1)$$

where v and $\overline{v}$ are two arbitrary videos (represented by the histograms H^i and $\overline{H}^i$, respectively), and

$$d(H^i, \overline{H}^i) = \sum_{b=1}^{B} \frac{(h^i_b - \overline{h}^i_b)^2}{h^i_b + \overline{h}^i_b}, \qquad (2)$$

is the χ^2 distance, whose average value is A_i for channel i.

In our work, we used only two channels (HOG and HOF) and the kernel (1) fed a support vector machine with default parameter $C = 10$ in a one-subject-out testing procedure[1].

We experimentally compared the STIP [17] and dense trajectories [27] on our dataset. For this purpose, we employ the available public codes[2] using the default parameters configuration for both representations. In the bag-of-feature encoding, different vocabulary sizes $\mathcal{S} = 600, 1000, 2000$ have been used for both STIP and dense trajectories. As the latter approach gives much more features (on average, 7588 per video against 428 for STIP), for the dense trajectories we also used $\mathcal{S} = 5000, 10000$ visual words (see Table 2 and 3). In addition, to limit the computational costs, we randomly sampled a subset of 900,000 HOG/HOF dense trajectory features to build the vocabulary.

[1] We compute seventeen accuracies training our systems on all the subjects except the one we are testing; then we average all the accuracies to get the final classification result.

[2] `http://lear.inrialpes.fr/software`

Comparisons	$L = 15$	$L = 5$
Pouring vs. Placing	82.76	**86.99**
Pouring vs. Drinking	70.50	**75.73**
Pouring vs. Passing	**74.79**	72.36
Passing vs Drinking	66.62	**67.13**
Passing vs Placing	**68.46**	67.58
Drinking vs Placing	68.39	**75.11**
All 4 intentions	46.76	**50.55**

Table 4. Trajectories shortening, $\mathcal{S} = 1000$

In Tables 2 and 3 we present the classification results comparing STIP and dense trajectories. Despite we changed the size $\mathcal{S}$ of the dictionaries (one for HOG and HOF, separately), we registered no significant variations in accuracy. Globally, the dense trajectories outperforms STIP in all the comparisons. Even though the IPs from STIP are supposed to be most descriptive in each frame, dense trajectories give much more features which led to a big gain in performance. Indeed, comparing STIP and dense trajectories with $\mathcal{S} = 1000$, the average improvement in the pairwise comparisons is +6.24% and +7.69% in all four comparisons. Thus, from now on, we will focus on dense trajectories.

5.2. Dense trajectory shortening

In order to tackle this extremely low variability between grasping motions, we tried to shorten the trajectory length changing the value of L from 15 to 5. With shorter trajectories, we tried to perform IfM analysis at a more atomic level. The rationale beside is that, since the clues that anticipate the intentions are almost invisible to the human eye (Section 3), analysing shorter motion patterns may allow us to predict future intentions by a collection of subtle movements more easily discriminated. Operatively, we switched to

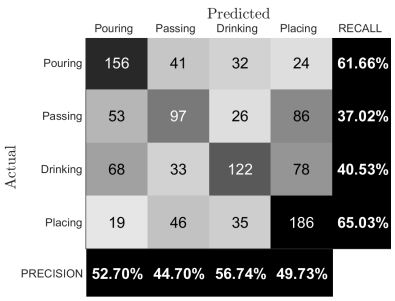


Figure 3. Dense trajectories $L = 5$ confusion matrix.

Comparisons	40%	50%	60%	70%	80%	90%	100%
Pouring vs. Placing	56,46	63,42	70,03	78,72	83,03	84,91	**86.99**
Pouring vs. Drinking	52,99	60,73	64,06	64,55	67,61	72,13	**75.73**
Pouring vs. Passing	61,02	61,77	62,57	65,05	65,55	69,83	**72.36**
Passing vs. Drinking	60,92	62,67	66,49	69,30	66,69	63,97	**67.13**
Passing vs. Placing	55,97	55,22	59,81	63,38	64,61	66,19	**67.58**
Drinking vs. Placing	57,78	58,87	61,38	64,80	69,29	73,82	**75.11**
All 4 intentions	31,79	34,58	38,41	41,16	42,81	46,52	**50.55**

Table 5. Accuracy percentage for the snippet analysis.

$L = 5$ and set $n_t = 1$; we also fixed $\mathcal{S} = 1000$ and, as done in Section 5.1, we sampled a subset of 900,000 HOG/HOF. The risk of this approach is that features may lose discriminative power: in a 25 fps acquisition, our $L = 5$ trajectories last $\frac{1}{5}$ seconds and, maybe, such time is too short to observe any useful motion cue. Fortunately, we found that this is not the case and, in fact, shorter trajectory features bring more evidence of the future intention. Actually, improving the beginning of the pipeline process results in a boost in the classification, showing that finer motion patterns are responsible of the actual intention. In Table 4, dense trajectory shortening is effective for IfM since provides +3.79% accuracy improvement in the all-class comparison.

5.3. Snippet analysis

Following the analysis in the previous section, we carried out a further investigation involving only small fragments of the video data. Shortening dense trajectories gives us the possibility of evaluating how much the movement is discriminant from the very beginning of the reach-to-grasp motion. In our case this is a further issue, since our dataset is made of very short videos, and considering only small portions we deal with only a bunch of frames for classification (e.g. in the 40% analysis, only 6 frames for the shortest video). The snippet analysis has been performed in the following way. We considered the whole set of HOG an HOF descriptors and we built a global bag-of-features dictionary with $\mathcal{S} = 1000$ words extracted from the 100% of the frames. Then, for the analysis at a fixed rate, we keep only the descriptor computed over any spatio-temporal cuboid completely included in the considered portion of the video. This requirement explains why we skipped 10%, 20% and 30%. Indeed, the shortest video lasts 16 frames and, thus, there are no spatio-temporal cuboids until we cover the 40% of that particular reach-to-grasp movement. In general, the accuracy increases as long as the percentage of the video considered grows (see Table 5): this means that bag-of-features histograms become more discriminative. Consequently, we can not find any portion of the reach-to-grasp execution that is useless for the prediction of intentions.

5.4. Discussion

Despite in Section 5.1 we showed that STIP [17] are less effective than dense trajectories [27], all the classification results in Table 2 always overcomes the random chance discrimination, with the remarkable 81.22% in Pouring vs. Placing task. Moving to the dense trajectory with $L = 5$ (Table 4), we have a high increase in performance on the pairwise comparisons (on average, +8.9% improvement), and, again, the easiest one is Pouring vs. Placing reaching 86.99%. Instead, Drinking and Passing are the most problematic intentions to recognize (see Figure 3). In Table 4, the all 4 intentions comparison leads to more impressive results: we indeed doubled the random chance level, scoring a 50.55%, and thus improving STIP by +11.48%. Consequently, we can conclude that, using computer vision, IfM motion problem is affordable in the sense that there exists some discriminative motion patterns that changes the execution of the reach-to-grasp movement, *de facto* anticipating the future intention. Moreover, the snippet analysis shows remarkable results already using very few frames of the videos (Section 5.3). In addition, such results are extremely valuable if considering the improvement with respect to human beings (Section 3).

Finally, one-subject-out testing procedure is suitable to devise an actual intention prediction system, dealing with human beings never seen before. However, it is much more demanding than a classical cross validation. For example, using $L = 5$ dense trajectory, with a 10-fold cross validation procedure for testing, we reach 85.43% in the all 4 comparison and 93.87% (on average) in the pairwise comparisons.

6. Conclusions

In this paper, we investigate the novel problem of Intention from Motion, consisting in the prediction of human intentions in a context-free setting, by only leveraging on kinematics. While proposing a novel multimodal dataset, we are able to show that the quantitative cognitive approach which rely on 3D motion capture data can be alternatively replaced with a video based paradigm where RGB optical videos are exploited. Despite the two data modalities are

actually very different, we register the following common trends. First, random chance is always exceeded: therefore a context-free intention prediction is actually feasible. Second, in terms of performance, we register a similar level of classification accuracies while using classical kinematic features from cognitive literature and state-of-the-art video-based approaches in computer vision. We demonstrate that exploiting RGB videos for intention prediction is 1) easier to acquire and 2) equally reliable for a quantitative analysis as its motion capture counterpart. This finding opens to a joint interdisciplinary approach in taking advantage of computer vision methods while tackling cognitive problems such as the prediction of human intentions.

References

[1] J. K. Aggarwal and M. S. Ryoo. Human activity analysis: A review. *ACM Computing Surveys*, 43(3), 2011.

[2] C. Ansuini, A. Cavallo, C. Bertone, and C. Becchio. Intentions in the Brain: The Unveiling of Mister Hyde. *The Neuroscientist*, 21:126–135, 2014.

[3] C. Ansuini, A. Cavallo, C. Bertone, and C. Becchio. The visible face of intention: why kinematics matters. *Front Psychol*, 5, 2014.

[4] C. Ansuini, A. Cavallo, A. Koul, M. Jacono, Y. Yang, and C. Becchio. Predicting object size from hand kinematics: a temporal perspective. *PloS One*, 2(10), 2015.

[5] D. N. Bub, M. E. J. Masson, and T. Lin. Features of planned hand actions influence identification of graspable objects. *Psychological Science*, 24(7):1269–1276, 2013.

[6] Y. Cao, D. Barrett, A. Barbu, N. Siddharth, H. Yu, A. Michaux, Y. Lin, S. Dickinson, J. M. Siskind, and S. Wang. Recognizing human activities from partially observed videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2013.

[7] I. Carpinella, J. Jonsdottir, and M. Ferrarin. Multi-finger coordination in healthy subjects and stroke patients: a mathematic modelling approach. *J Neuroeng Rehabil*, 8(1), 2011.

[8] A. Cavallo, A. Koul, C. Ansuini, F. Capozzi, and C. Becchio. Decoding intentions from movement kinematics. In *Scientific Reports*, 2016.

[9] A. Chakraborty and K. Roy-Chowdhury. Context-aware activity forecasting. In *Asian Conference on Computer Vision (ACCV)*, 2014.

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

[11] D. F. Fouhey and C. Zitnick. Predicting object dynamics in scenes. In *IEEE Computer Vision and Pattern Recognition (CVPR)*, 2014.

[12] M. Hoai and F. De la Torre. Max-margin early event detectors. In *IEEE Computer Vision and Pattern Recognition (CVPR)*, 2012.

[13] J. M. Kilner. More than one pathway to action understanding. *Trends in Cognitive Sciences*, 15:352–357, 2011.

[14] K. M. Kitani, B. D. Ziebart, J. A. Bagnell, and M. Hebert. Activity forecasting. In *European Conference on Computer Vision (ECCV)*, 2012.

[15] H. Koppula and A. Saxena. Anticipating human activities using object affordances for reactive robotic response. In *Robot Sci Syst*, 2013.

[16] T. Lan, T.-C. Chen, and S. Savarese. A hierarchical representation for future action prediction. In *European Conference on Computer Vision (ECCV)*, pages 689–704, 2014.

[17] I. Laptev. On space-time interest points. *Int. J. Comput. Vision*, 64(2-3):107–123, 2005.

[18] V. Manera, C. Becchio, A. Cavallo, L. Sartori, and U. Castiello. Cooperation or competition? Discriminating between social intentions by observing prehensile movements. *Experimental Brain Research*, 211(3-4):547–556, 2011.

[19] M. Pei, Y. Jia, and S.-C. Zhu. Parsing video events with goal inference and intent prediction. In *IEEE International Conference on Computer Vision (ICCV)*, 2011.

[20] M. S. Ryoo. Human activity prediction: Early recognition of ongoing activities from streaming videos. In *IEEE International Conference on Computer Vision (ICCV)*, 2011.

[21] M. S. Ryoo, T. J. Fuchs, L. Xia, J. K. Aggarwal, and L. Matthies. Robot-centric activity prediction from first-person videos: What will they do to me? In *International Conference on Human-Robot Interaction (HRI)*, 2015.

[22] K. Schindler and L. J. V. Gool. Action snippets: How many frames does human action recognition require? In *IEEE Computer Vision and Pattern Recognition (CVPR)*, 2008.

[23] I. Tsochantaridis, T. Joachims, T. Hofmann, and Y. Altun. Large margin methods for structured and interdependent output variables. *J. Mach. Learn. Res.*, 6:1453–1484, Dec. 2005.

[24] M. van Elk. The left ipl represents stored hand-postures for object use and action prediction. *Frontiers in Psychology*, 5(333), 2014.

[25] C. Vondrick, H. Pirsiavash, and A. Torralba. Anticipating the future by watching unlabeled video. *CoRR*, abs/1504.08023, 2015.

[26] J. Walker, A. Gupta, and M. Hebert. Patch to the future: Unsupervised visual prediction. In *IEEE Computer Vision and Pattern Recognition (CVPR)*, 2014.

[27] H. Wang, A. Kläser, C. Schmid, and L. Cheng-Lin. Action Recognition by Dense Trajectories. In *CVPR*, 2011.

[28] D. Xie, S. Todorovic, and S.-C. Zhu. Inferring "dark matter" and "dark energy" from videos. In *IEEE International Conference on Computer Vision (ICCV)*, 2013.

[29] K. Yamaguchi, A. C. Berg, L. E. Ortiz, and T. L. Berg. Who are you with and where are you going? In *IEEE Computer Vision and Pattern Recognition (CVPR)*, 2011.

[30] G. Yu, J. Yuan, and Z. Liu. Predicting human activities using spatio-temporal structure of interest points. In *ACM Multimedia Conference*, 2012.

[31] M. Ziaeefard and R. Bergevin. Semantic human activity recognition: A literature review. *Pattern Recognition*, 48(8):2329 – 2345, 2015.