Intention prediction approach to interact naturally with the microworld

Laura Cohen, Sinan Haliyo, Mohamed Chetouani, Stéphane Régnier

Intention prediction approach to interact naturally with the microworld

Laura Cohen, Sinan Haliyo, Mohamed Chetouani and Stéphane Régnier

Abstract— **Micromanipulation tools are not yet commonly used in the industry or in the research due to the lack of natural and intuitive human-computer interfaces. This work proposes a vision based approach using a Kinect RGB-Depth sensor to provide a "metaphor-free" interface. An intention prediction approach is proposed, based on a cognitive science computational model, to allow a more natural interaction without any prior instructions. This model is compared to a vision based gesture recognition approach in terms of naturalness and intuitiveness. It shows an improvement in user performance in terms of duration and success of the task, and a qualitative preference for the proposed approach evaluated by a user survey.**

I. INTRODUCTION

Manipulation of micro objects is a key issue for further developments in nanotechnology. At this scale, typical manipulation tasks involve picking and placing micro objects. They cannot be handled manually due to their reduced size, complex and counter-intuitive force fields, and the high influence of environmental conditions [1] [2]. Furthermore, the vision feedback and sensor data of a micromanipulation systems are generally limited. Non-repetitive and unautomated tasks therefore imply the design of dedicated interfaces between the user and the micro world. Recent works focus on the development of natural and intuitive interfaces, to provide information from the microworld toward the natural human sensory modalities. Hence, haptic interfaces enable the operator to touch these intangible scales [3], and virtual reality to see and interact with the microworld in 3D [4].

Despite the recent progress, the lack of intuitive interfaces limits the adoption of micromanipulation systems. From the operator's point of view, these interfaces remain complex to use, unhandy and hard to manage. For example, naive users mention the difficulty to identify the haptic arm handle and the real end-effector. Furthermore, these interfaces require a learning phase because of the use of a specific symbolic language. Several works in the literature agree on the necessity to create more natural and intuitive interfaces for widespread adoption of microrobotic technologies, but this specific issue has not been addressed yet from the user perspective.

New solutions are currently emerging in the field of macroscale Human Computer Interface (HCI), with low-cost RGB-Depth sensors (e.g. Microsoft Kinect) based on natural interaction modalities such as hand gesture recognition, without requiring to wear markers or gloves. This kind of interfaces, dedicated to macroscale interaction, appears also as a promising acquisition solution for micromanipulation interfaces as it allows high intuitivity and benefits from robust detection tools like skeleton tracking. These sensors are widely used in video games for interaction tasks that involve the whole body, like dance or tennis games whereas typical microscale interaction consists in more precise manipulation tasks like picking, displacing and placing objects. Current macroscale solutions widely use gesture recognition methods to detect predefined gestural commands. Ren and O'Neill [5] propose a 3D selection method that uses the gesture direction to determine the target object and confirm selection by raising the other hand. This kind of interfaces therefore requires to learn and remember a complex metaphorical language [6]. Furthermore, their low flexibility makes them poorly robust to gestures that differ from the predefined dictionary. Even if the gesture modality seems "natural", the need to produce a discrete predefined gesture remains as metaphorical and symbolic as clicking on a button to trigger an action. It implies a learning phase and thus is ill-fitted for nonspecialist users.

To address this issue, "metaphor-free" interfaces have been proposed [7]. The aim of this approach is to enable the user to interact with virtual objects in the same way as they interact with the real world. Hence, the operator relies on already mastered skills and doesn't need any instruction. This approach raises many challenges. An interface that would need no instructions to specify the constraints of the system to the user has to handle the complexity of free interaction. A metaphor-free interface thus requires interpreting the natural human behavior. Humans are experts in understanding others' behavior. A promising approach is to exploit models issued from cognitive science research on how humans infer others' intentions by observing their actions.

The objective of this work is to provide a natural interface to assist the virtual teleoperation of an AFM (Atomic Force Microscope) probe for micromanipulation. User movements are tracked using a Kinect sensor to allow free interaction. A metaphor-free intention prediction model is proposed to grab and drop the microsphere, based on cognitive science findings in the field of intention prediction. This approach is compared to a classical hand gesture recognition system in terms of naturalness and intuitiveness.

II. AFM VIRTUAL TELEOPERATION SYSTEM

At microscale, typical interactive manipulation tasks consist in picking, transfering and placing micro objects on a substrate. An example strategy of micromanipulation that has been experimentally demonstrated for this purpose is the pick-and-place by adhesion with an AFM cantilever [8] [9].

Authors are with Institut des Systèmes Intelligents et de Robotique, Université Pierre et Marie Curie - Paris 6, CNRS UMR 7222, 4 Place Jussieu, 75005 Paris, France. laura.cohen@isir.upmc.fr

A. AFM based micromanipulation

For objects with dimensions less than $100\mu m$, adhesion forces (namely van der Waals, capillary and electrostatic) are stronger than gravitation forces. These forces can be advantageously used for manipulation. An AFM cantilever is used for the manipulation as shown on fig. 1. Objects are picked by simple contact due to the superiority of the adhesion vs. weight. Then, objects are placed on a substrate with an higher surface adhesion than the cantilever's.

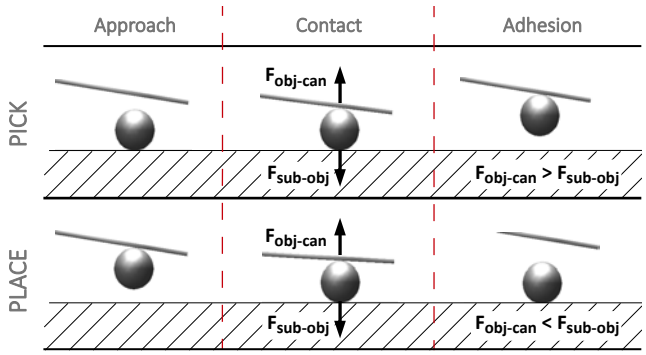

Fig. 1. Pick (up) and place (down) of a microsphere with an AFM cantilever by adhesion. F_{i-j} is the adhesion force between i and j.

An AFM based micromanipulation interface hence has to fulfill the following subtasks :

- picking an object by adhesion with the AFM cantilever
- transfering the object while avoiding involuntary contact with the substrate
- placing the object on a target site by adhesion with another substrate

B. Teleoperation of an AFM cantilever

Some recent works have focused on providing both haptic and 3D virtual reality feedbacks to enable the operator to manipulate micro objects intuitively in a wide range of applications [10] [11]. Virtual reality simulators are interesting solutions to provide users with a reconstructed 3D view of a manipulation scene, which can be displayed at the human scale and from different vantage points.

In a previous work [12], authors used an haptic arm to teleoperate an AFM cantilever in a virtual reality simulator. This simulator realistically reproduces the adhesion and dynamics of a real system and validates haptic teleoperation. It is composed of a flexible virtual cantilever that can be displaced depending on the system kinematics. The adhesion forces between the object, the substrate and the cantilever shown on fig. 1 are realistically modeled. The interaction is measured from the cantilever deflection as in a real AFM system. A physics engine enables real-time interaction. The haptic feedback allows the user to feel micro interactions like adhesion and pull-off phenomenons.

The same simulator is used here to explore new ways to interact with the microworld that could provide an assistance for virtual teleoperation. Instead of the haptic device, a Kinect sensor is used for acquisition of users' motions. These RGB-Depth sensors are promising solutions, with less constraints for the user and specially dedicated to natural 3D interaction.

C. Virtual reality simulator

To allow a direct link between the user's hand and the virtual effector, this work proposes to add a natural interface layer in the simulator. In this interface, the user interacts through a virtual hand included in the simulator. The Kinect skeleton joints of the operator are used to estimate the hand's 3D position and rotation in the virtual reality environment, providing a direct way to interact with the microworld. The proposed virtual teleoperation system is shown in fig. 2. The virtual hand's position is then used to teleoperate the realistic simulation with a master-slave system.

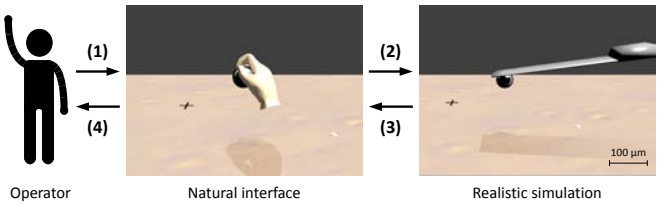

Fig. 2. Teleoperation of the micromanipulation simulation. (1) The user's hand translation and rotation are used to move the virtual hand in the natural interface. (2) Then, a master-slave system is used to teleoperate the AFM cantilever in the simulation. (3) The failure/success of the pick/place task is fed back towards the natural interface to give visual feedback to the operator (4).

In this work, two approaches are compared to detect the grabbing and dropping actions performed by the operator, associated to the pick-and-place task of the microsphere. For each task, corresponding animations of the virtual hand are created to provide visual feedback to the operator. When a gesture is recognized, the corresponding animation is played and the grab/drop of the microsphere is triggered. For this purpose, a gesture recognition method is presented in the next section and is compared to a new intention prediction approach, detailed in section IV. These two approaches are compared and validated using the AFM manipulation simulator.

III. Hand gesture recognition

To detect the grabbing and dropping actions, a classical method is to use gesture recognition. In order to avoid the constraint of learning metaphorical gestures, a first approach is to use "natural" gestures. As the proposed task involves the pick-and-place of a microsphere, two natural gestures are used : a closed hand near a sphere triggers the pick and an open hand near a target triggers the drop. The OpenNI [13] hand gesture recognition library is used as it is a robust and position-invariant method.

A preliminary evaluation of this system consists in the observation of users interacting without any instruction on the method to pick-and-place the sphere. A first issue is

that the animation can be triggered only when the gesture is recognized, that is when the gesture is completed. Therefore, users report an impression of delay between the performed gesture and the virtual hand animation, even if the gesture recognition operates in real-time. The second issue is that users tend to not fully close their hand, as if to adapt it to the size of the virtual object to be grabbed. Furthermore, the rest position of the hand is neither totally open nor closed. Despite the robustness of the hand gesture recognition method, many false negatives are observed. The interface seems sufficiently close to a real interaction for the user to behave as if to interact with a real object, however, this likeness harms the robustness of the detection.

There is therefore a duality between the natural behavior of the user towards the interface and the symbolic language used by it. This observation shows that it is not sufficient to mimic a natural behavior, like grabbing/dropping an object by associating to it a hand open/closed recognition. To address this issue, a metaphor-free intention prediction model is proposed in the next section. The gesture recognition approach is used in this work as a baseline to evaluate this intention prediction model.

IV. Computational model of intention prediction

The observations from the previous section show that designing a natural interface requires fitting to the natural behavior of the user. In order to perform the task, two pieces of information are required: what object the user wants to interact with, and what action to perform on this object. An action is composed of two parts: an intention and a movement. In cases of deliberative action, the action is caused by what Searle [14] terms a prior intention, that is, an intention to act formed in advance of the action itself. A same movement can be caused by different prior intentions, for example a person can grasp an apple to eat it or to hand it to another person. Becchio et al. [15] show that human observers can infer the prior intention of others by observing their actions. In this paper, this capacity is termed as intention prediction.

Building a computational model of intention prediction is an interesting perspective for further developments of HCI. In fact, intention prediction approach can lead to an earlier detection than classical gesture recognition, thus avoiding the delay between the gesture recognition and the visual feedback. Intention is a high level notion and thus cannot be directly extracted by a sensor. It is however based on low level data that are quantifiable. In this work, a high level intention prediction model is proposed based on low level features extracted from the Kinect sensor.

In order to establish a computational model of intention prediction, the first step consists in determining the relevant low level quantifiable data upon which to model the intention of the user.

A. Low level features to model the intention

Despite the great variability of human goal-directed gestures, there exist some invariant features. The velocity has an invariant log normal shape, that can be approximated by a Gaussian for the movements with an intermediate speed [16]. This classical profile is shown in fig. 3 on the left from a gesture towards a target in the virtual reality simulator described in the previous section. Another property of goal-directed gestures is the isochrony principle [17]. The duration of a gesture to reach a target is demonstrated to be a constant, i.e. the speed depends linearly on the distance to the target. This result is validated in the proposed architecture from 60 trials of reaching a target in the virtual reality simulator as shown in fig. 3 on the right.

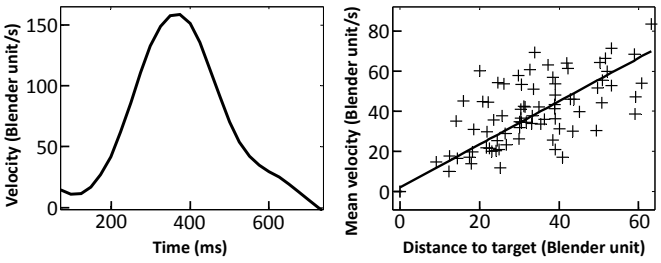

Fig. 3. Invariant velocity profile for goal-directed gesture and isochrony principle. A Blender unit corresponds to the Blender simulator default unit of measurement.

Several works in the field of cognitive science have studied intention prediction. Becchio et al. [15] synthetize various experimental data showing that for a same gesture, for example a prehension gesture, the kinematics features vary depending on the prior intention of the actor. Sartori et al. [18] demonstrate that kinematics features are sufficient for an observer to infer the prior intention. Gesture kinematics contains both invariant features and sufficient information to enable a human observer to infer others' intentions. Thus, the kinematic features are selected as low level information upon which to build the intention prediction model. For this purpose, hand gesture velocity is extracted by derivating the position of the hand joint of the Kinect SDK skeleton library.

B. High level model of intention prediction

Despite the fact that humans are experts at intention prediction, cognitive models of this capacity are not exploited to design natural interfaces. Oztop et al. [19] propose a computational model that could give an explanation of how the same cerebral areas are used in motor planning and in understanding others' action. Based on an intention hypothesis, the observer predicts the sensory consequences according to his own learned motor behavior. At the next step, a difference is computed between the prediction and the observation. If the prediction error is too high, a new intention hypothesis is then made. This model is predictive and metaphor-free, and therefore is an interesting solution to answer the raised issues. This work proposes to adapt this model to predict the intention of a user in the context of the micromanipulation interface. The proposed system architecture is shown in fig. 4.

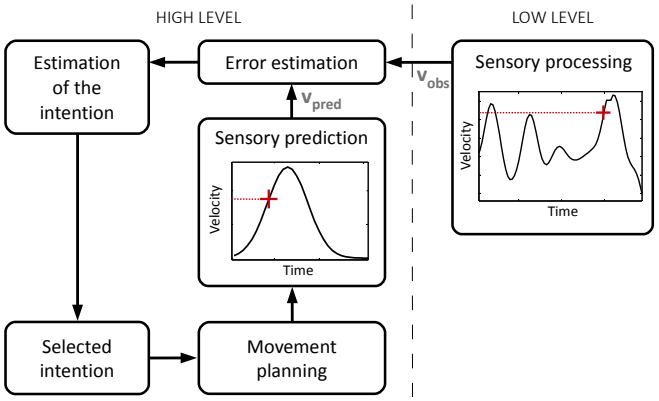

Fig. 4. Intention prediction cognitive model

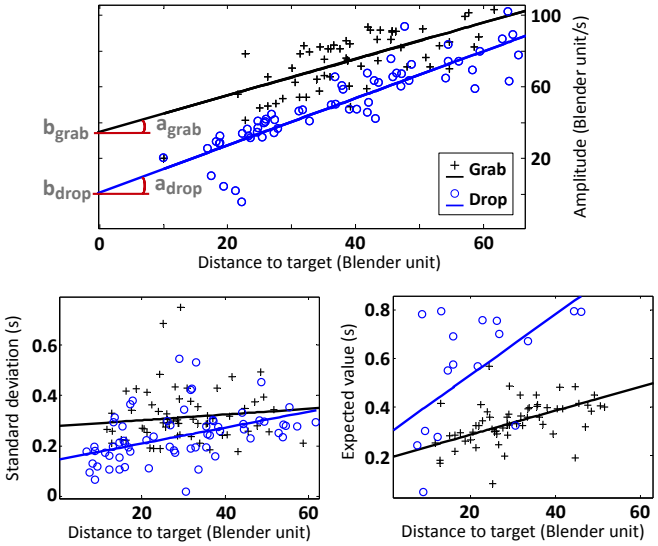

Fig. 5. Task influence on the Gaussian predictor parameters

C. Construction of predictors

The first step is to learn the basic motor behavior of grab and drop gestures upon which to compute the intention predictors. The predictor profile is approximated by a Gaussian according to the invariant speed profile law. The isochrony principle states that gesture velocity increases with the initial distance to the target. Thus, the amplitude of the Gaussian should be proportional to the distance to the target.

The learning phase protocol consists in 20 grab gestures and 20 drops at various distances from the target performed by 3 subjects. A microsphere and a target dropping site are randomly placed on the substrate for each round. The grabbing and dropping are automatically triggered when the hand is sufficiently close to the target, in order to avoid any a priori hypothesis that could affect the naturalness of the interaction. No information is given to the user about how to grab and drop the object. A Gaussian fit is then performed on the hand velocity for each task to compute the amplitude, expected value and standard deviation parameters. An example result for a user learning phase is shown on fig. 5. The three parameters (amplitude (A), expected value (EV) and standard deviation (SD)) of the Gaussian are plotted as functions of the initial distance to the target. Each point corresponds to one round (in black for the grab task and in blue for the drop task). A polynomial interpolation is then performed on the points for each parameter and each task.

As expected, the amplitude depends linearly on the distance to the target. Only the first order of the polynomial interpolation is significant. This result is consistent with the isochrony law. Furthermore, the expected value is also a linear function of the distance to the target. On the other hand, the standard deviation doesn't seem to depend on the distance to the target. Another interesting result is that the curves associated with the grab and the drop tasks are different. Despite the fact that both are goal-directed gestures, the amplitude and the expected value of the speed vary depending on the intention. This observation supports the fact that prior intention modifies the movement kinematics even for the same reach gesture [15] [18]. Thus, different predictors need to be computed for these actions. The predictors $pred(t)$ are computed based on these observations according to the current possible task $\{grab, drop\}$ and to the current distance to target ($dist$), based on equation (1), where SD_{task} and EV_{task} are the means of respectively the standard deviation and expected value learned for each possible task on 3 users, and $A_{dist,task}$ is the amplitude of the Gaussian predictor computed from linear equation (2).

$$pred(t) = A_{dist,task} \cdot \mathrm{e}\left(-\frac{(t - EV_{task})^2}{2 \cdot SD_{task}^2}\right) \quad (1)$$

$$A_{dist,task} = a_{task} \cdot dist + b_{task} \quad (2)$$

a_{task} and b_{task} are parameters learned from the linear interpolation of the amplitude in function of the distance to the target, depending on the grab or drop task, as shown on fig. 5. As the standard deviation doesn't seem to depend on the distance to target, its value is computed by a mean on the learning samples. Furthermore, the gesture segmentation is emergent from the proposed model. A gesture starts when it is predicted well enough. Thus, the expected value is simply evaluated as a mean on the learning samples.

D. Prediction method

In the first phase of the task, the user grabs the microsphere. According to the proposed model, the intention hypothesis is set on the grab intention, and the corresponding predictor is used. Algorithm 1 presents the proposed prediction method for the grab task. At each step, if the prediction mean error is under a fixed threshold ($errThresh$), the predictor is considered as good enough and it is kept. The score (N) is increased. An intention hypothesis is validated if the predictor has been good for $Npred$ points. This threshold is fixed depending on the frame rate of the sensor, shortly after the maximum of the Gaussian is reached. Once the grab

intention is predicted, the intention hypothesis is set to the drop task and the following steps are repeated.

Algorithm 1 Grab intention prediction

Extract current distance to target $dist(t)$ from sensor
$velocity(t) = \frac{d(dist(t))}{dt}$
if $velocity(t) > 0$ **then**
  $\overline{err} = \frac{1}{N} \cdot \sum_{i=1}^{N} (pred(N) - velocity(t))^2$
  if $\overline{err} > errThresh$ **then**
    $pred(t) = A_{dist,grab} \cdot \mathrm{e}\left(\frac{(t-EV_{grab})^2}{2 \cdot SD_{grab}^2}\right)$
    $N = 1$
  else
    $N = N + 1$
  end if
end if
if $N > Npred$ **then**
  Grab intention predicted
end if

Fig. 6 shows an example of a grabbing task intention prediction. The first 2s correspond to non-goal-directed random gestures. The figure shows that the predictor is not good enough, therefore many new predictors are recalculated. After 2s, the gesture corresponds to a reach to grasp the microsphere. Hence, the mean error between the predictor and the user mean hand velocity ($\overline{err}$) is bellow the threshold $errThresh$ and after $NPred$ points, the grab intention is predicted, corresponding to the red dotted line. Once

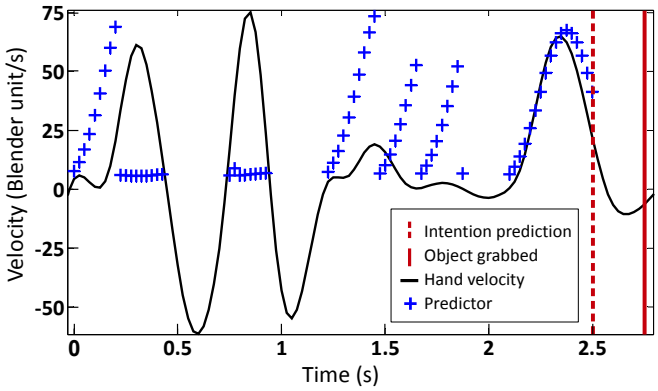

Fig. 6. Intention prediction example

the microsphere grab intention is predicted, the interface starts the corresponding animation of the virtual hand, in order to avoid the delay observed with the classical gesture recognition method.

V. Experimental results

A. User test protocol

A user test protocol is set up to evaluate the proposed interactive system and compare it to the baseline hand gesture recognition approach. The task consists in grabbing a microsphere and displacing it to a target on the substrate marked as a cross. A new configuration of the microsphere and the target is drawn randomly at the end of each round. To validate the natural aspect of the interaction, no instruction is given to the operator on the method to use to perform the task. The experimental protocol is conducted in 3 phases. The first one is a case-control in which the grabbing and dropping is triggered only by a proximity criterion as in the learning phase. The second uses the gesture recognition method presented in section III. The third uses the intention prediction approach to test the model proposed in this work. The order of the phases proposed to the user are changed randomly in order to avoid a learning effect on the results. Each phase is constituted of 15 rounds of grab and drop. Nine adult subjects took part in the experiment. At the end of each phase, a user questionnaire is submitted to the operator. This questionnaire is based on the System Usability Scale (SUS) [20].

B. Comparative results on intention prediction and gesture recognition

1) Quantitative results: The success of the task is evaluated on a proximity and duration criterion : if the hand stays near the target for more than $0.5s$ and the grab/drop is not detected, the task is considered as a failure. The success percentage for each task is shown in fig. 7 on the left. The proposed intention prediction method shows a significant improvement of 33.8% over the classical gesture recognition method for the dropping task. A second quantitative result

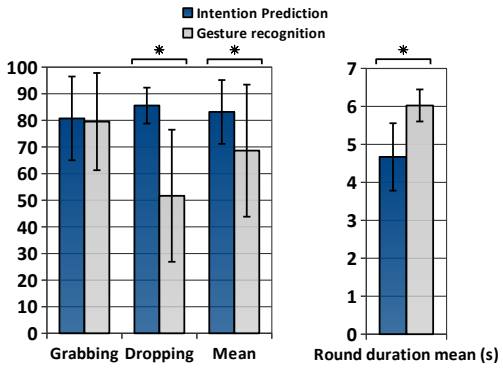

Fig. 7. Percentage of success for the two tasks with the intention prediction and classical gesture recognition approaches (left) and round duration mean (right). One asterisk (*) indicates a p value smaller than 0.05 ($p < 0.05$). The statistical analysis used is the ANOVA test.

is the mean duration of each round. The intention prediction methods mean duration is $4.7s$, while the gesture recognition is $6s$ as shown in fig. 7 on the right. Thus, the proposed method allows for significant improvement in the duration of the task.

2) Qualitative results: Finally, the SUS questionnaire shows a strong preference for the model proposed in this work of 30.8% over the gesture recognition method as shown in fig. 8. In particular, the users' evaluation of the ease to manipulate the microspheres shows better results than both the hand gesture recognition method and the case-control method. The evaluation of comfort is consistent with these

results, showing a significant preference for the proposed approach.

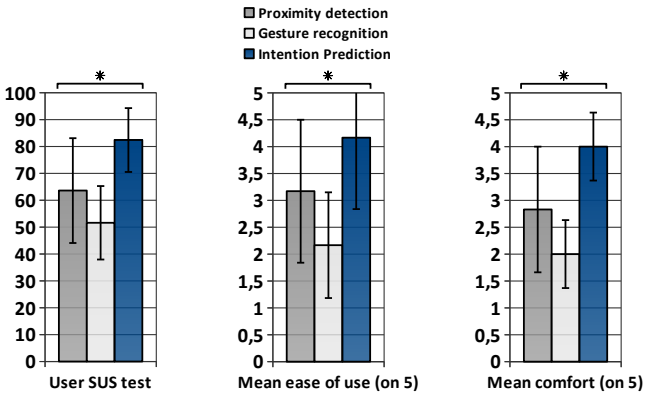

Fig. 8. Qualitative SUS user survey results. One asterisk (*) indicates a p value smaller than 0.05 ($p < 0.05$). The statistical analysis used is the ANOVA test.

C. Discussion

The results show the improvements of the proposed method in a context of natural interaction without any prior instruction, both quantitatively and qualitatively. Hence, the proposed approach is validated as an appropriate solution to provide an assistance for AFM micromanipulation. However, some users mention the fact that the method seems sometimes too predictive, weakening the impression of being in control of the task.

VI. CONCLUSIONS

In this paper, a metaphor-free interface is proposed and applied to the teleoperation of an AFM cantilever for a virtual drag and drop task of microspheres. To achieve a non symbolic interaction, an intention prediction model is proposed, based on the natural invariant features of human goal-directed movement. Without any instruction to users, the predictivity of this model significantly improves the task duration over the classical hand gesture recognition method. In addition, the SUS questionnaire shows a 30.8% preference of the users for the intention prediction model over the hand gesture recognition approach. Hence, these results confirm the metaphor-free approach as an interesting solution for natural interaction. Furthermore, the proposed approach enables manipulation tasks independently of the hand pose and the fingers. Thus, this method can be generalized to various systems, such as a mouse or a haptic arm. For further improvements, the operator's focus of attention can be integrated to generalize the model to multitarget contexts. For this purpose, an attentional module selects the more pertinent target depending on the focus of attention of the user.

REFERENCES

[1] M. Gauthier and S. Régnier, *Robotic Micro-assembly*. IEEE press, 2010.

[2] S. Régnier and N. Chaillet, *Microrobotics for Micromanipulation*. Wiley-ISTE, 2010.

[3] A. Bolopion, H. Xie, D. S. Haliyo, and S. Régnier, "Haptic teleoperation for 3-d microassembly of spherical objects," *Mechatronics, IEEE/ASME Transactions on*, vol. 17, no. 1, p. 116127, 2012.

[4] B. Sauvet, N. Ouarti, S. Haliyo, and S. Régnier, "Virtual reality backend for operator controlled nanomanipulation," in *3M-NANO, 2012 International Conference on*, p. 121127, 2012.

[5] G. Ren and E. O'Neill, "3D selection with freehand gesture," *Computers & Graphics*, vol. 37, pp. 101–120, May 2013.

[6] D. A. Norman, "Natural user interfaces are not natural," *Interactions*, vol. 17, p. 610, May 2010.

[7] P. Ravasio and V. Tscherter, "Users theories of the desktop metaphor, or why we should seek metaphor-free interfaces," *Beyond the desktop metaphor: designing integrated digital work environments*, p. 265294, 2007.

[8] D. S. Haliyo, S. Régnier, and J.-C. Guinot, "MuMAD, the adhesion based dynamic micro-manipulator," *European Journal of Mechanics-A/Solids*, vol. 22, no. 6, pp. 903–916, 2003.

[9] F. Dionnet, D. S. Haliyo, and S. Régnier, "Autonomous micromanipulation using a new strategy of accurate release by rolling," in *Robotics and Automation, 2004. Proceedings. ICRA'04. 2004 IEEE International Conference on*, vol. 5, pp. 5019–5024, IEEE, 2004.

[10] G. Millet, A. Lécuyer, J.-M. Burkhardt, S. Haliyo, and S. Régnier, "Haptics and graphic analogies for the understanding of atomic force microscopy," *Int. J. Hum.-Comput. Stud.*, vol. 71, p. 608626, May 2013.

[11] A. Bolopion, C. Stolle, R. Tunnell, S. Haliyo, S. Régnier, and S. Fatikow, "Remote microscale teleoperation through virtual reality and haptic feedback," in *IROS, 2011 IEEE/RSJ International Conference on*, pp. 894–900, IEEE, 2011.

[12] G. Millet, A. Lécuyer, J.-M. Burkhardt, D. S. Haliyo, and S. Régnier, "Improving perception and understanding of nanoscale phenomena using haptics and visual analogy," in *Haptics: Perception, Devices and Scenarios* (M. Ferre, ed.), no. 5024 in Lecture Notes in Computer Science, pp. 847–856, Springer Berlin Heidelberg, Jan. 2008.

[13] http://www.openni.org/.

[14] J. R. Searle, *Intentionality: An essay in the philosophy of mind*. Cambridge University Press, 1983.

[15] C. Becchio, V. Manera, L. Sartori, A. Cavallo, and U. Castiello, "Grasping intentions: from thought experiments to empirical evidence," *Frontiers in human neuroscience*, vol. 6, 2012.

[16] H. Nagasaki, "Asymmetric velocity and acceleration profiles of human arm movements," *Experimental Brain Research*, vol. 74, pp. 319–326, Jan. 1989.

[17] P. Viviani and T. Flash, "Minimum-jerk, two-thirds power law, and isochrony: converging approaches to movement planning," *Journal of Experimental Psychology: Human Perception and Performance*, vol. 21, no. 1, pp. 32–53, 1995.

[18] L. Sartori, C. Becchio, and U. Castiello, "Cues to intention: the role of movement information," *Cognition*, vol. 119, no. 2, p. 242252, 2011.

[19] E. Oztop, D. Wolpert, and M. Kawato, "Mental state inference using visual control parameters," *Cognitive Brain Research*, vol. 22, pp. 129–151, Feb. 2005.

[20] J. Brooke, "SUS-A quick and dirty usability scale," *Usability evaluation in industry*, vol. 189, p. 194, 1996.