frontiers in Robotics and AI

ORIGINAL RESEARCH
published: 07 June 2017
doi: 10.3389/frobt.2017.00014

Detecting Biological Motion for Human–Robot Interaction: A Link between Perception and Action

Alessia Vignolo[1,2], Nicoletta Noceti[1], Francesco Rea[2], Alessandra Sciutti[2], Francesca Odone[2] and Giulio Sandini[1]*

[1] Dipartimento di Informatica, Bioingegneria, Robotica e Ingegneria dei Sistemi (DIBRIS), Università degli Studi di Genova, Genova, Italy, [2] Robotics, Brain and Cognitive Science Department (RBCS), Istituto Italiano di Tecnologia, Genova, Italy

One of the fundamental skills supporting safe and comfortable interaction between humans is their capability to understand intuitively each other's actions and intentions. At the basis of this ability is a special-purpose visual processing that human brain has developed to comprehend human motion. Among the first "building blocks" enabling the bootstrapping of such visual processing is the ability to detect movements performed by biological agents in the scene, a skill mastered by human babies in the first days of their life. In this paper, we present a computational model based on the assumption that such visual ability must be based on local low-level visual motion features, which are independent of shape, such as the configuration of the body and perspective. Moreover, we implement it on the humanoid robot iCub, embedding it into a software architecture that leverages the regularities of biological motion also to control robot attention and oculomotor behaviors. In essence, we put forth a model in which the regularities of biological motion link perception and action enabling a robotic agent to follow a human-inspired sensory-motor behavior. We posit that this choice facilitates mutual understanding and goal prediction during collaboration, increasing the pleasantness and safety of the interaction.

Keywords: biological motion, two-thirds power law, temporal multi-resolution motion descriptor, HRI, robot attention

Edited by:
Trung Dung Ngo, University of Prince Edward Island, Canada

Reviewed by:
Tung Xuan Truong, Le Quy Don Technical University
Fady Alnajjar, RIKEN Brain Science Institute (BSI), Japan
Raffaella Lanzarotti, University of Milan

****Correspondence:***
Alessia Vignolo
alessia.vignolo@iit.it

Specialty section:
This article was submitted to Vision Systems Theory, Tools and Applications, a section of the journal Frontiers in Robotics and AI

Received: *24 November 2016*
Accepted: *24 April 2017*
Published: *07 June 2017*

Citation:
Vignolo A, Noceti N, Rea F, Sciutti A, Odone F and Sandini G (2017) Detecting Biological Motion for Human–Robot Interaction: A Link between Perception and Action. Front. Robot. AI 4:14. doi: 10.3389/frobt.2017.00014

1. INTRODUCTION

Robots are progressively entering our houses: robotic devices as vacuum cleaners, pool cleaners, and lawn mowers are becoming more and more commonly used and the growth of robotics in the consumer sector is expected to continuously increase in the near future.[1] The fields of applications for robotics will influence not only domestic activities but also entertainment, education, monitoring, security, and assistive living, leading robots to frequent interactions with untrained humans in unstructured environments. The success of the integration of robots in our everyday life is then subordinated to the acceptance of these novel tools by the population. The level of comfort and safety experienced by the users during the interaction plays a fundamental role in this process. A key challenge in current robotics has then become to maximize the naturalness of human–robot interaction (HRI), to foster a pleasant collaboration with potential non-expert users. To this aim, a promising avenue seems to be endowing robots with a certain degree of social intelligence, to enable them to behave appropriately in human environments.

[1] EU Strategic Road Map 2014–2020.

In this context, human infants represent an important source of inspiration. Indeed, even if endowed with limited sensory-motor capabilities and no explicit knowledge of social norms, young children can already quite proficiently coordinate with their peers (Asendorpf and Baudonniere, 1993) and caregivers (Tomasello et al., 2005), even in the absence of language. Moreover, from the restricted social abilities exhibited in the very first months of life, humans are able to develop a full-fledged social competence in adulthood. The partial skills exhibited by a baby can, therefore, represent the minimum set of abilities necessary to enable the bootstrapping of more complex interactive expertise. Endowing robots with analogous "social building blocks" represents, therefore, the starting point in the attempt to replicate complex HRI skills, favoring the establishment of a simple yet efficient intuitive understanding in the naive user.

In this work, we consider in particular the marked, natural predisposition of newborns to notice potential interacting partners in their surroundings, which is manifested by a preference for biological motion (Simion et al., 2008) and for faces looking directly to them (Farroni et al., 2002) over other visual stimuli. Interaction in its simplest form seems, therefore, constituted by a sensitivity to some properties of others' motion and to their direction of attention.

Drawing inspiration from these observations, we propose a video-based computational method for biological motion detection, which we also implement on the humanoid robot iCub (Metta et al., 2010a), to guide robot attention toward potential interacting partners in the scene. We focus on a method purely based on motion, which does not require any *a priori* knowledge of human shape or skeleton, nor detecting faces and hands (Brethes et al., 2004; Gaschler et al., 2012).

In essence, we put forth a model in which the regularities of biological motion link perception and action enabling a robotic agent to follow a human-inspired sensory-motor behavior. This way, we address two fundamental components necessary to facilitate the understanding of robots by human users:

1. On the *perception* side, we make the robot *find the same types of stimuli salient* as a human (Breazeal and Scassellati, 1999). In particular, we propose a computational tool to make the robot sensitive to human activity, a very relevant type of motion for human observers.
2. On the *action* side, we enable the robot to *direct its attention to human activity* through a biologically inspired oculomotor mechanism (Breazeal et al., 2001). This way the robot can reorient its gaze toward where the human partners are acting. Such eye shift can also represent an intuitive form of communication, revealing where the robot is focusing and potentially informing the human partner of its availability to interact (Palinko et al., 2015).

The use of a common, biologically inspired, perceptual, and motor framework facilitates the human partner's understanding and prediction of the future actions of its robot counterpart.

To design a system sensitive to the regularities typical of biological movements we draw inspiration from the laws governing human motor control. We consider in particular the *Two-Thirds Power Law*, since there is an evidence that human neonates are sensitive to it since the first days after birth (Méary et al., 2007). The law is a well-known invariant of human movements (Viviani and Stucchi, 1992; Viviani et al., 1997; Vieilledent et al., 2001; Richardson and Flash, 2002) describing the regular relationship between the instantaneous tangential velocity and the radius of curvature of human end point movements (Greene, 1972; Viviani and Terzuolo, 1982; Lacquaniti and Terzuolo, 1983). There have been experimental evidence, particularly for handwriting (Viviani and Terzuolo, 1982; Lacquaniti et al., 1983), that in biological movements velocity and curvature show a strong mutual influence. The low-level motion descriptor we adopt, based on the same dynamic features, is meant to capture such connections.

To handle the wide intra-class variability of biological stimuli, we propose the use of a structured motion descriptor that accounts for multiple temporal resolutions of the measurements. A careful, automatic selection of such resolutions allows us to easily adapt our model to a variety of scenarios.

The method, preliminarily discussed in the study of Vignolo et al. (2016a) is here deeply investigated by using a much wider set of combination of temporal windows to filter the features. We also test the method on a much wider set of variations including different sensors, point of view, type of behaviors, and dynamics. In particular, its efficacy in generalizing to new scenarios, including scene observation from different visual perspectives and the presence of severe occlusions, is demonstrated.

The possibility to exploit such a method for robot perception is then validated by implementing the method in a module integrated in the software framework of the iCub humanoid robot (Vignolo et al., 2016b). The module implements an engineered variation of our method—appropriately handled to work online and in real-time—and is used to enhance the robot visual attention system, endowing the robots with the ability to rapidly redeploy attention on actions performed by human agents in the scene with a biologically plausible saccadic behavior. The advantage of the solution is that attention is biased toward moving human agents even when they are not visible in the scene. At the same time, the natural robot gaze motion can act as an implicit communication signal, informing the collaborators of its current attentional state. In this paper, a detailed analysis of the results of the integration between the motion classification and the attentive system is done by separating the two stages of perception and action, giving a better idea of when, during the robot pipeline, and why the robot fails or is not perfectly precise in the discrimination task between the biological and non-biological movements. Moreover, we also analyze the velocity profile of the fixation point to reach the target.

The rest of the paper is organized as follows. In the next section, we provide a review of works related to our approach. Section 3 is devoted to present in detail our method, followed by an extensive experimental analysis discussed in Section 4. The iCub architectural framework that hosts our method on the robot is presented in Section 5, while in the following section we show both the results produced on the method while working online and the effect on the robot action. Section 7 is finally left to a discussion on possible future outcomes.

2. RELATED WORKS

Several are the approaches that have been adopted so far to *perceive* and *detect* human activities. In the following, we discuss a tentative state-of-art, while enhancing the novelties of our approach.

One potential approach to detect humans is to endow robots of specific sensors such as RFID or thermal sensors (Correa et al., 2012). In spite of high performances, this solution requires *ad hoc* hardware, usually not available in common robotics platform, limiting the range of possible scenarios. Their relatively high cost is another factor that may harm a large-scale diffusion—which is, however, desirable for future family or companion robots.

For these reasons, we focus here on approaches based on more traditional RGB and depth sensors. Although there is a large number of works in the computer vision field, the constraints and limitations of robotics setting make it difficult to directly employ methods successfully applied to other domains. Robots are in general equipped with relatively low-resolution cameras, in order not to overload their network, while standard computer vision approaches may rely on high-resolution images. Moreover, interactive robots require a fast processing to support interaction: a perfect classification performance becomes useless if it is achieved not rapidly enough to enable appropriate robot reaction. In this respect, the speed-precision trade-off in HRI is often unbalanced toward speed, as a rapid, yet not precise estimation still allows the robot to continue the collaboration, while adjustment of the initial guess may always be achieved exploiting the evolution of the interaction itself.

With these constraints in mind, we cite here examples of use of RGBD sensors (Sung et al., 2012; Koppula et al., 2013) promoted in recent years by the widespread availability of low-cost, highly portable sensors. This approach provides a richer information on the body structure, helping the understanding of the performed activity, but to the price of low success when the visibility of the partner is limited and it is not possible to match the 3D input with the human skeleton.

More related to our work is a third category of approaches, based on the analysis of 2D video signals acquired with the robot cameras (Dillmann, 2004; Weinland et al., 2011).

Most 2D video analysis methods for human detection currently adopted in robotics rely on appearance or shape features, for instance detecting faces and hands in the scene (Brethes et al., 2004; Gaschler et al., 2012). However, these approaches have severe limitations as scene complexity grow, for instance when the clutter in the environment increases or the light conditions become more challenging. Shape-based or part-based methods are likely to fail when the human body is only partially visible—as in presence of occlusions—while detectors based on faces are not appropriate for close interaction scenarios, as those involving precise manipulation on a tabletop. Although still based on 2D signals, our approach substantially differs from previous works in that we strictly focus only on the motion properties of the stimuli. A purely motion-based human detection system makes it possible to detect the presence of humans in the vicinity just by observing the effects of their behavior on the environment, as for instance, the movement of the manipulated tool—a use case that to the best of our knowledge has not been considered so far in the related literature. Note that, while motion detection is common in robotics applications, oftentimes as a preliminary step for further analysis, *human detection through motion* requires a selectivity to biological motion, which is usually absent in common robotic systems.

There is wide evidence that humans are better at predicting stimuli moving according to biological motion, whereas they present a distorted perception when behaviors subvert these kinematics rules (Viviani et al., 1997; Pozzo et al., 2006; Elsner et al., 2012; Gavazzi et al., 2013). Also in the specific context of HRI, it has been demonstrated that the adoption of biological plausible motion by a humanoid robot can lead to a more natural coordination with its actions (Bisio et al., 2014) and potentially to a more pleasant interaction (Sciutti et al., 2012). Conversely, the execution of non-biological motion by a humanoid robot has been suggested as a possible cause for the Uncanny Valley effect (Mori, 1970), i.e., to the occurrence of a sense of eeriness and disgust toward the robot, precluding the possibility for a natural interaction (Chaminade and Cheng, 2009). Human-like motion benefits interaction also when it is applied to gaze behavior, for instance facilitating the regulation of conversations (Mutlu et al., 2012), the coordination of shared plans in collaboration (Boucher et al., 2012) and the prediction of robot goals (Rea et al., 2016). Drawing inspiration from these evidence, to maximize the efficacy of the human activity detection module, our proposed architecture leverages the regularities of biological motion also for the preparation and execution of the robot saccadic action. This way, the robotic oculomotor action triggered by the perception module informs the human partner in an intuitive way about the internal attentional status of the robot.

3. A TEMPORAL MULTI-RESOLUTION BIOLOGICAL MOTION DESCRIPTOR

In this section, we start with a brief summary of an instantaneous motion description we adopt as a building block for our method (Noceti et al., 2015b). Then, we review the proposed multi-resolution method (Vignolo et al., 2016a), which efficiently combines measurements that may span different temporal portions of an image sequence.

3.1. Instantaneous Motion Representation

We report in **Figure 1** the key steps of our low-level layer of motion representation. At each time instant t, the *optical flow* is computed using a dense approach (Farnebäck, 2003), which provides an estimate of the apparent motion vector in each image point (**Figure 1B**). The optical flow magnitude is thresholded to enhance locations with significant motion. Isolated pixels and small regions, which are likely to be generated by noise, are rejected by first applying a *perceptual grouping*—in which only locations whose neighboring pixels are also marked as moving are kept in the analysis—and then discarding small groups. We then obtain a motion map whose largest connected component (henceforth referred to as $\mathcal{R}(t)$) becomes the candidate region for

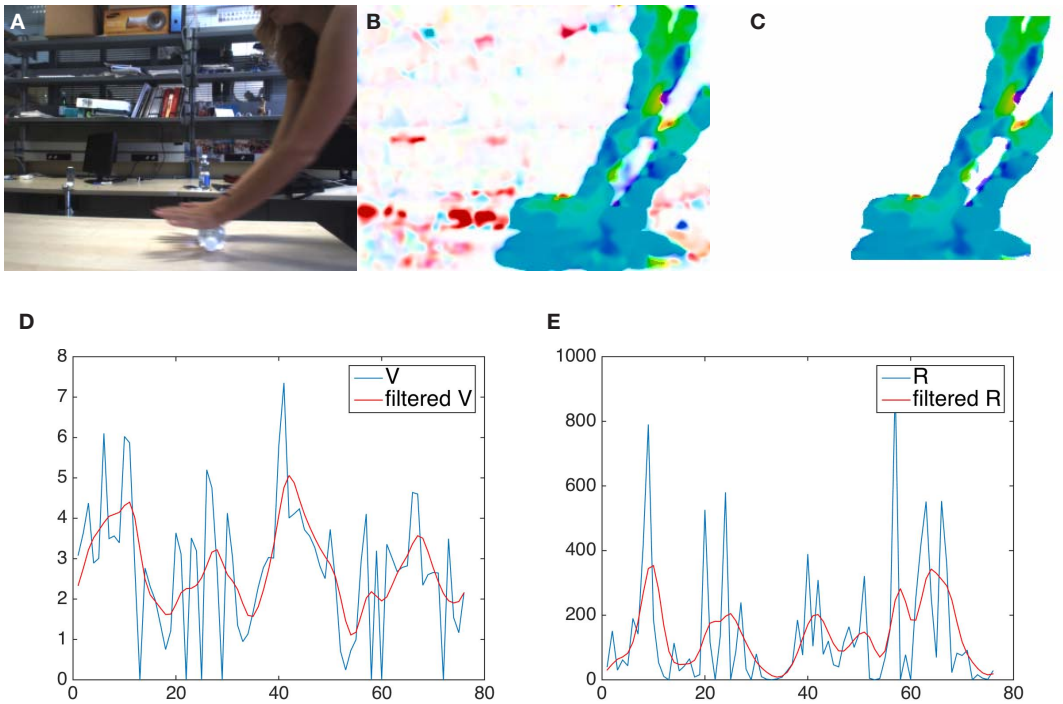

FIGURE 1 | The key steps of our low-level motion representation. First row, from left: **(A)** the original frame, **(B)** the map of the optical flow magnitude and **(C)** the segmented region. Below: a visual comparison between raw (blue) and filtered (red) features; **(D)** velocity and **(E)** radius of curvature.

TABLE 1 | Empirical formulations of the spatio-temporal dynamic features (Δ_t is the temporal displacement between observations of two adjacent time instants).

Tangential velocity	$\hat{V}_i(t) = (u_i(t), v_i(t), \Delta_t)$
Tangential velocity magnitude	$\hat{V}_i(t) = \sqrt{(u_i(t)^2 + v_i(t)^2 + \Delta_t^2}$
Acceleration	$\hat{A}_i(t) = (u_i(t) - u_i(t-1),\ v_i(t) - v_i(t-1), 0)$
Curvature	$\hat{C}_i(t) = \frac{\|\hat{V}_i(t) \times \hat{A}_i(t)\|}{\|\hat{V}_i(t)\|^3}$
Radius of curvature	$\hat{R}_i(t) = \frac{1}{\hat{C}_i(t)}$
Angular velocity	$\hat{A}_i(t) = \frac{\hat{V}_i(t)}{\hat{R}_i(t)}$

motion recognition (**Figure 1C**), under the assumption that only a single interesting source of motion is observed in the scene at each time instant.

Let $(u_i(t),v_i(t))$ be the optical flow components associated with point $\mathbf{p}_i(t) \in \mathcal{R}(t)$, and N the size of the region, i.e., the number of pixels in it. We compute a set of motion features, according to the formulations in **Table 1**, which empirically estimate the analytical quantities related by the Two-Thirds Power Law. We finally describe the region $\mathcal{R}(t)$ with a feature vector $x_t \in \mathbb{R}^4$ by averaging the features over all the region elements:

$$\mathbf{x}_t = \frac{1}{N}\left[\sum_i \hat{V}_i(t), \sum_i \hat{C}_i(t), \sum_i \hat{R}_i(t), \sum_i \hat{A}_i(t)\right] \quad (1)$$

Figure 1, on the bottom line, shows the behavior of two of the computed features (velocity and radius of curvature) across a period of time lasting 80 frames. As expected, the peculiarities of the performed movements are best appreciated by observing it for some time.

3.2. Multi-Resolution Motion Representation over Time

Since a *meaningful event* lasts more than one temporal tick, we may integrate the instantaneous motion representation over a fixed temporal frame w. To this purpose, we consider a set of w subsequent measurements $[\mathbf{x}_{t-w}, \ldots, \mathbf{x}_t]$ and compute a running average of each feature across time, obtaining a new motion descriptor $\hat{\mathbf{x}}_t(w)$.

The choice of an appropriate size for the temporal window is critical and highly dependent on the specific dynamic event. For this reason, we adopt a multi-resolution approach, where different temporal windows are jointly adopted, and we propose an adaptive procedure where we learn from examples the best combination of temporal windows.

More in detail, let us consider a maximum temporal window extent $w_T^{MAX} \in \mathbb{N}$, such that $w_T^{MAX} > 1$, and a selection of potentially interesting time windows w_T defined as elements of a set $W = \left\{w \in \mathbb{N} | w \geq 1 \wedge w \leq w_T^{MAX}\right\}$.

At a certain time instant t we may have a temporal sequence of observations $S_t \in \mathbb{R}^{4w_T^{MAX}}$ as

$$S_t = \left[\mathbf{x}_{t-w_T^{MAX}}, \ldots, \mathbf{x}_t\right]. \quad (2)$$

We apply a bank of *running average* filters—of widths selected from the range in the set W—to each feature separately. The result

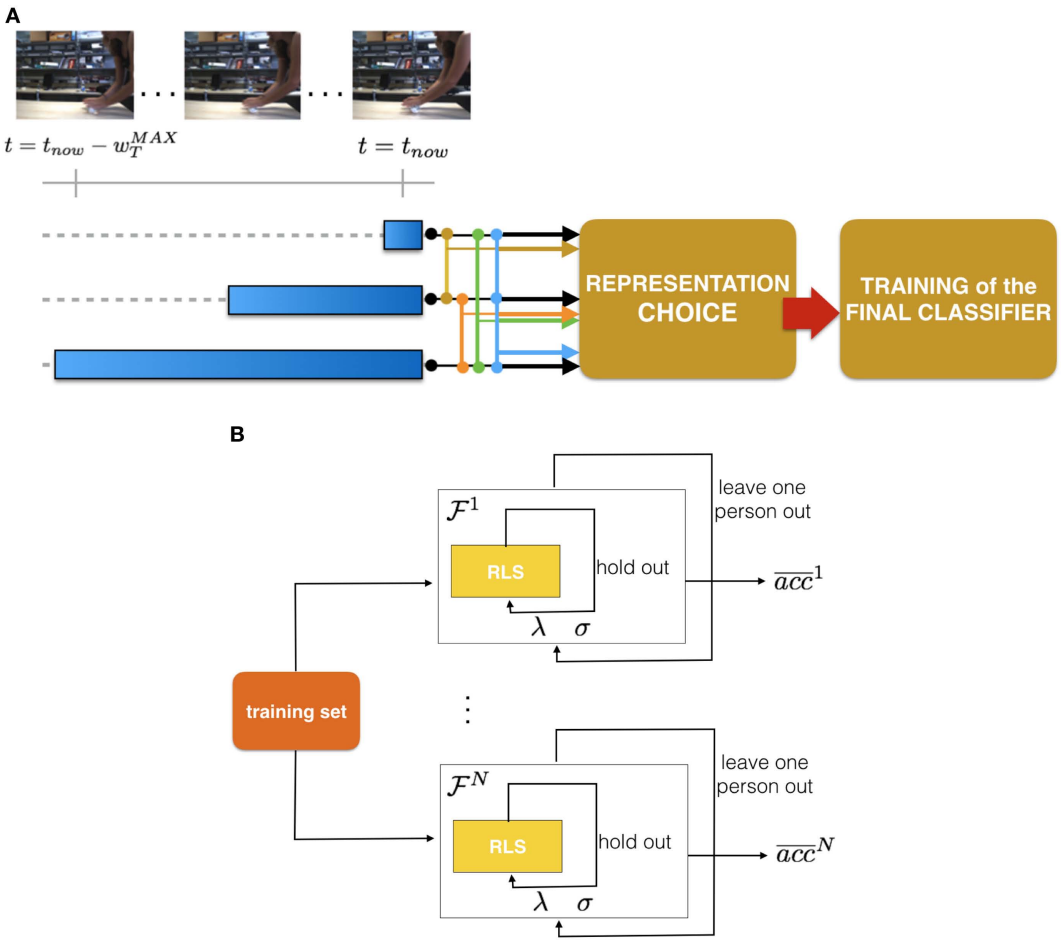

FIGURE 2 | (A) A visual sketch of our pipeline. From left: in each image of a sequence we detect the moving region and compute the features. We then compute a running average of those features over different temporal windows (3 in this example, identified by the blue ranges). Then, we evaluate different temporal models which may be composed of more than one temporal resolution and select the most appropriate. **(B)** A visual sketch representing the core procedure of our model selection.

is a set of motion descriptors $\hat{\mathbf{x}}_t(w_T)$ referring to different time periods w_T and such that

$$\begin{aligned} \hat{\mathbf{x}}_t(1) &= \mathbf{x}_t \\ \hat{\mathbf{x}}_t(w_T) &= \mathcal{RA}(S_t|_{w_T}, w_T), \text{ for } 1 < w_T \leq w_T^{MAX} \end{aligned} \tag{3}$$

where $\mathcal{RA}$ is the running average filtering while the notation $S_t|_{w_T}$ denotes the restriction of sequence S_t to the last w_T elements. The leftmost part of **Figure 2A** reports a sketch of this filtering procedure.

Starting from the set of motion descriptors of Eq. 3, we obtain many possible temporal multi-resolution motion descriptors $\{\mathcal{F}_t^i\}_i$:

$$\mathcal{F}_t^i = \oplus \delta^i(w_T)\hat{\mathbf{x}}_t(w_T), \text{ for all } w_T \in W \tag{4}$$

where $\oplus$ denotes the concatenation between feature vectors, while $\delta^i(w_T) \in \{0, 1\}$ is a binary weight representing the presence or absence of the corresponding filtered vector in the final descriptor.

Thus, as a final step, we need to select an appropriate and minimal combination of different temporal windows, considering that a multi-resolution descriptor will allow us to deal with different types of dynamic events, but many different temporal windows would carry a similar amount of information. The core of the selection process is detailed in the next section, as it is intertwined with the actual motion classification step.

3.3. Biological Motion Representation and Classification

We formulate the problem of recognizing biological motion from video sequences as a binary classification problem. To this purpose, given a certain temporal scheme denoted with $i^\star$, we consider a training set.

$$Z = \{(\mathcal{F}_k^{i^*}, y_k) \in X \times Y\}_{k=1}^n \tag{5}$$

where $\mathcal{F}_k^{i^*} \in X \subseteq \mathbb{R}^d$ is a given temporal multi-resolution descriptor (input),[2] while $y_k \in Y = \{-1, 1\}$ is the associated output label (1 for biological samples and -1 for negative non-biological

[2]We omit the index t of Eq. 4 for readability.

samples). The size d depends on the specific $\mathcal{F}_k^{i^*}$ considered. Henceforth, we will refer to $\mathcal{F}_k^{i^*}$ as $\mathcal{F}_k$.

To learn the relationship between input and output in a predictive way, we adopt a *Regularized Least Squares* (RLS) binary classifier which amounts at minimizing the following functional.

$$f_Z = \arg\min_{f\in\mathcal{H}} \frac{1}{n}\sum_{k=1}^{n}(y_k - f(\mathcal{F}_k))^2 + \lambda||f||_{\mathcal{H}} \tag{6}$$

where $\mathcal{H}$ is a Reproducing Kernel Hilbert Space with a positive semi-definite kernel function K, and λ a regularization parameter that controls the trade-off between the data term and the smoothness term. At run time, a new datum $\mathcal{F}$ is associated with an estimated label obtained by the sign of $f_Z^{\lambda}(\mathcal{F})$, with

$$f_Z^{\lambda}(\mathcal{F}) = \sum_{k=1}^{n}\alpha_k K(\mathcal{F}, \mathcal{F}_k) \tag{7}$$

where $\alpha = (\mathbf{K} + n\lambda I)^{-1}\mathbf{y}$ is an n-dimensional vector of unknowns, while $\mathbf{K}$ is the associated kernel matrix. In the model selection procedure better detailed in Section 4.1.2, we train a set of classifiers each one associated with a different combination of motion features $\mathcal{F}$. The best multi-resolution motion descriptor is selected in a data-driven manner, by ranking the validation error achieved by the different classifiers.

4. OFFLINE EXPERIMENTAL ANALYSIS

In this section, we discuss the experiments we performed primarily on video sequences acquired with the iCub humanoid robot (Metta et al., 2010a), using the machine learning library GURLS for an efficient implementation of RLS (Tacchetti et al., 2013). Our classifier is equipped with an RBF kernel, following our conclusions in the study of Vignolo et al. (2016a). To evaluate the sensitivity of the method to the acquisition sensor, we also considered test sets captured with a common webcam and a hand-held camera (Canon EOS 550D).

In the following, we first discuss in detail the training procedure. Second, we show the generalization capability of our approach by discussing its appropriateness on a selection of tests including new dynamic events, new scenarios, and on data acquired by a different sensor.

4.1. Training the Motion Classifier

The training phase of a motion classifier includes *(i)* a *model selection* in which the classification parameters and the most appropriate *multi-resolution representation* are chosen; and *(ii)* training of the final classifier based on the previously selected model.

4.1.1. The Data Set

Our training set is composed of indoor videos of three subjects observed by the iCub eyes while performing repetitions of given actions from a repertoire of dynamic movements typical of a human–robot interaction setting. The choice of acquiring a collection of videos in-house is due to the absence, to the best of our knowledge, of a benchmark explicitly designed for purposes similar to ours. More in detail, we consider *Rolling dough* (9 movements, ~300 frames—**Figure 3A**), *Pointing* a finger toward a certain 3D location (7 movements, ~330 frames—**Figure 3B**), *Mixing in a bowl* (29 movements, ~190 frames—**Figure 3C**), *Transporting* an object from and to different positions on a table (6 movements, ~300 frames—**Figure 3D**), and *Writing* on a paper sheet (3 movements, ~300 frames—**Figure 3E**). As for the non-biological examples, we consider a selection of dynamic events which can be observed indoors: a *Wheel with a random pattern* (~300 frames—**Figure 3F**) and a *Wheel with a zig-zag pattern* (~300 frames—**Figure 3G**), a *Balloon* (~300 frames—**Figure 3H**), a *Toy Top* turning on a table (~300 frames—**Figure 3I**), and a *Toy Train* (~398 frames—**Figure 3J**).

For each dynamic event we acquired two videos. Henceforth, we will adopt the notation $\{V_{S_i1}\}$ and $\{V_{S_i2}\}$, $i = 1, 2, 3$, to denote, respectively, the sets of first and second video instance of subject S_i. Similarly, $\{V_{N1}\}$ and $\{V_{N2}\}$ are the two sets of videos containing non-biological events.

In the following, the training set used for training the classifier and selecting the model includes $\{V_{S_i1}\}$ for $i = 1, 2, 3$, $\{V_{N1}\}$, and $\{V_{N2}\}$. Details on how they are divided are described where appropriate. Instead, $\{V_{S_i2}\}$, $i = 1, 2, 3$, are left out and used as a first test in Section 4.2. The images have size 320×240 and have been acquired at an approximate rate of 15 fps. The cameras we used in our work (both the robot and the opposite view webcam used for the test) have a relatively low resolution.

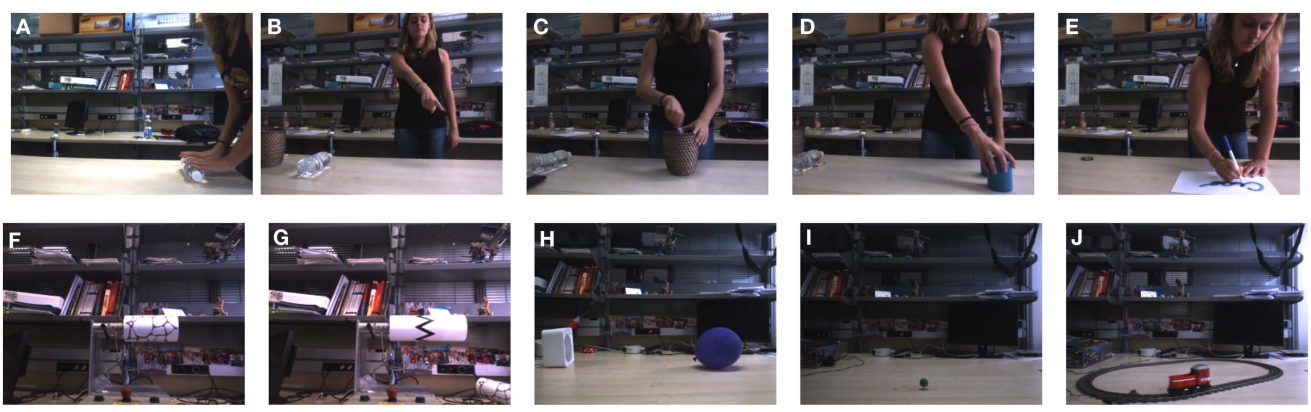

FIGURE 3 | Biological and non-biological movements included in the training set. **(A)** Rolling dough, **(B)** pointing, **(C)** mixing in a bowl, **(D)** transporting, **(E)** writing, **(F)** wheel (random pattern), **(G)** wheel (zig-zag) pattern, **(H)** balloon, **(I)** toy-top, and **(J)** toy train.

4.1.2. Model Selection

The main purpose of the model selection (see a visual sketch in **Figure 2B**) step is to choose the most appropriate temporal multi-resolution representation, from a large set of N choices. This will allow us, at run time, to compute that representation only.

We perform the selection in a data-driven manner, where for each representation considered, we obtain an average validation accuracy by adopting a *Leave-One-Subject-Out* approach.

4.1.2.1. Leave-One-Subject-Out Procedure

For a given multi-resolution representation (Eq. 4), we represent all data accordingly, then we partition the training set each time leaving the videos of one subject $\{V_{S_i1}\}$ as positive examples of a validation set. As for the negatives, the set $\{V_{N1}\}$ is always used as a training and the set $\{V_{N2}\}$ as a validation. This allows us to obtain an average validation accuracy. Notice that within each run of the training procedure we include a *hold out* process (with $M = 10$ different partitioning), with a balanced training, that allows us to select the parameters σ (RBF Kernel parameter) and λ (RLS regularization parameter).

4.1.2.2. Detecting the Best Representation

Let N be the number of the different multi-resolution representations considered. This number depends on the cardinality of the set of potentially interesting time windows W (see Section 3.2). We set $W = \{1, 5, 10, 15, 20, 25, 30\}$ and forced the final temporal descriptor [equation (4)] to be concatenation of at most 3 different temporal windows [equation (3)]. We chose a maximum size temporal window of 30 frames—equivalent to 2 s—as this temporal period already affords complex action processing in human brain (Urgen et al., 2012).

The step size of 5 frames between adjacent windows is due to the intrinsic nature of the data. The choice of considering at most 3 temporal windows is suggested by the need of controlling the amount of data redundancy. Under these assumptions, we obtain $N = \#W + \binom{\#W}{2} + \binom{\#W}{3} = 63$.

In **Figure 4A**, we show the performances of each representation scheme, ranked in descending order with respect to the average validation accuracy. The bars are color-coded with respect to the number of concatenated temporal representations (from dark to light: 3, 2, 1). In general, three temporal windows appear to be more descriptive, and in particular the ones including different temporal ranges (short–medium–long) are ranked first. As a single temporal window, the 30 frames choice performs on average very well.

With this analysis, we conclude that the temporal multi-resolution representation that concatenates the raw features vector with the filtered measures on temporal windows $w_T = 15$ and $w_T = 30$ is the best-performing, leading to a final feature vector of length 12. **Figure 4B** shows the classification accuracies of the selected multi-resolution representation, compared to the cases a single filter width is adopted, on the validation set. A first observation is that there is not a single temporal window appropriate for all the events: for instance, the single filter width $w_T = 30$ performs quite well in all cases but one (sequence *Mixing*, case (c) of **Figure 4B**), as the very fast dynamics of the movement requires smaller window sizes for filtering the signals. Indeed, shorter time windows provide better performances in this case.

Overall, the multi-resolution descriptor reports more stable performances, with higher average accuracies and lower SDs (see **Table 2**). This speaks in favor of the capability of our approach to cope effectively with dynamic events of variable temporal duration when no prior information is available.

4.1.3. Training the Final Classifier

Now we have selected the most appropriate temporal representation $r^\star$, we may build the final classifier. To this purpose, we consider the whole training set and run a final training procedure using the $r^\star$ representation (1-15-30), and performing model selection in order to set $\sigma^\star$ and $\lambda^\star$ again with a balanced hold out procedure, with $M = 10$ trials. The obtained classifier is adopted to evaluate the capability of our method to generalize to new data, as discussed in the following sections.

4.2. Testing the Classifier

In this section, we report the results of our testing analysis (see **Figure 5**). The experiments we carried out aim at testing the validity of the model on new data, including data containing very different appearance of dynamics with respect to the training set.

We organized the experiments in different test trials, to discuss the robustness of our model on scenarios of increasing complexity. At first, we perform an assessment of the method on the same actions of the training set but using different videos (Test I in the following). Then, we proceed considering conditions that vary with respect to the training set: we focus on movements included in the training set but characterized by different speeds or trajectory patterns (Test II); actions in critical situations of visibility, as in the presence of occlusions, limited spatial extent of the observed motion, and even when just the shadow is in the camera field of view (Test III); different human actions (on the fronto-parallel plane or performed in depth with respect to the camera) recorded with the robot (Test IV) and with a hand-held camera placed in front of the robot, to test the influence of the acquisition sensor and of the viewpoint (Test V).

In the following, we discuss in detail each test.

4.2.1. Positive Examples

We focus on biological movements and consider the training actions, adopting the second set of videos of each subject, i.e., $\{V_{S_i2}\}, \{V_{S_22}\}, \{V_{S_32}\}$. As expected, the method performs very well (see the graph in **Figure 5A**), with an average accuracy, across subject, of 0.98 ± 0.03.

4.2.2. Negative Examples with Changing Speed

We consider variations of the apparent motion with respect to the training set. Case 1: the three training subjects performing faster training actions *Rolling dough, Transporting*; Case 2: the *Wheel* (with one of the appearance patterns of the training set) and the *Toy train* (with the same trajectory of the training set) with different speed and the *Wheel* with a different pattern (first picture in **Figure 5B**); Case 3: the *Toy train* covering a circular trajectory (second picture in **Figure 5B**) as opposed to the ellipsoidal path considered in the training set (**Figure 3J**), with slower and faster

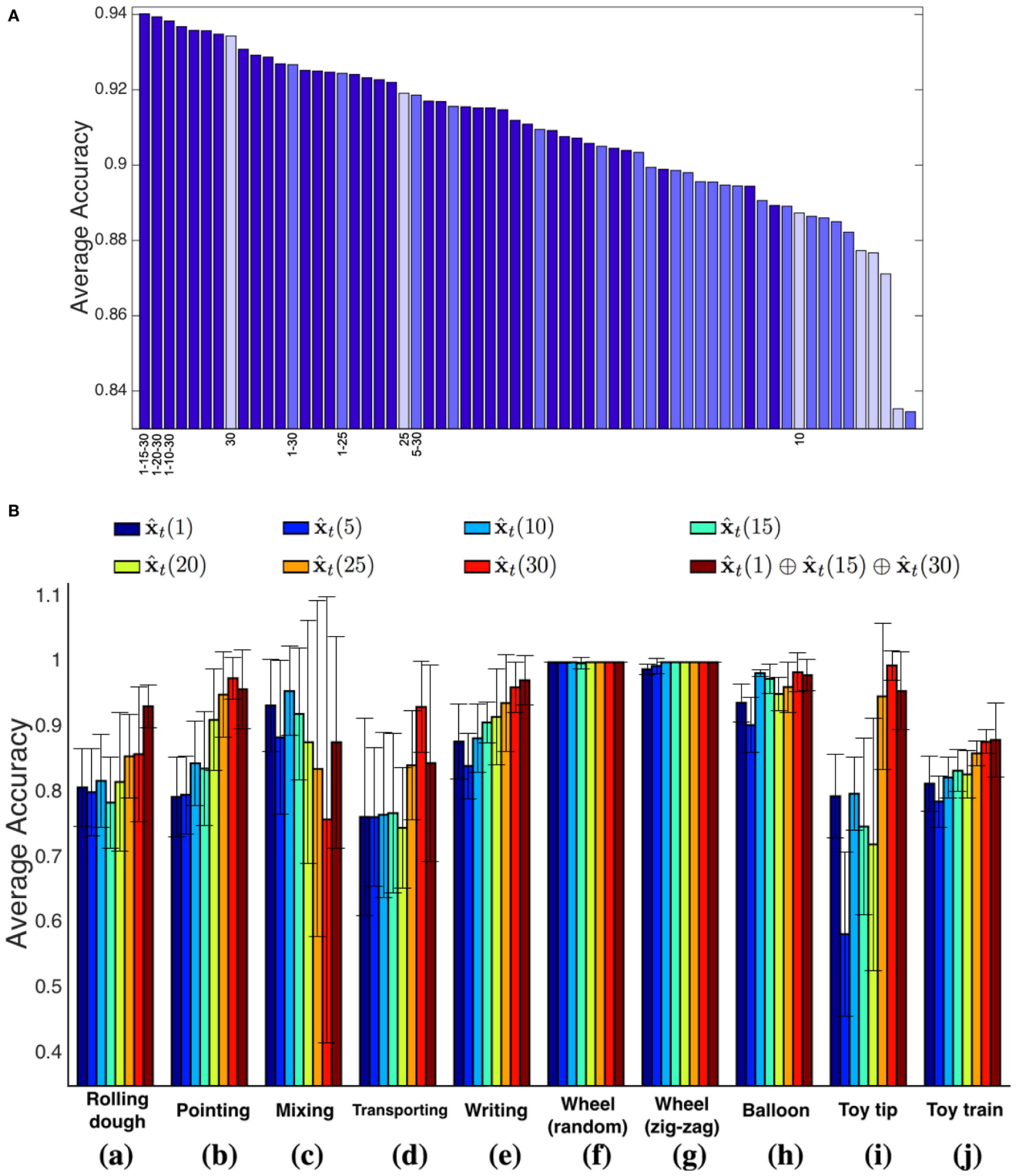

FIGURE 4 | (A) Average accuracy for each representation (dark blue: concatenation of three temporal windows; medium blue: concatenation of two temporal windows; light blue: a single temporal window). **(B)** A comparison between the selected temporal multi-resolution descriptor and all the different time widths considered independently. The results refer to the actions of **Figure 3**.

TABLE 2 | Average and SD accuracy of the temporal single-resolution representations and the best performing multi-resolution scheme.

Representation	Average accuracy	SD acc.
$\hat{x}_t\,(1)$	0.87	0.09
$\hat{x}_t\,(5)$	0.84	0.12
$\hat{x}_t\,(10)$	0.89	0.09
$\hat{x}_t\,(15)$	0.88	0.10
$\hat{x}_t\,(20)$	0.88	0.10
$\hat{x}_t\,(25)$	0.92	0.06
$\hat{x}_t\,(30)$	0.93	0.08
$\hat{x}_t(1) \oplus \hat{x}_t(15) \oplus \hat{x}_t(30)$	**0.94**	**0.05**

Bold font refers to the best accuracy we have obtained.

velocity profiles (at approximately half and twice the velocity of the training set). The accuracies, reported in the graph in **Figure 5B**, show again very appropriate values, although an influence of the variations applied in Case 3 can be observed. This may be explained with a partial lack on information when the conditions become too severe (presence of high velocity, limited spatial extent of the apparent moving region).

4.2.3. Occlusions and Distant Dynamics

We focus here on some critical scenarios.

- **Case 1**: a training subject performs actions included in the training set (see an example in the first picture in **Figure 5C**) and a new one (*Waving*) with partial occlusions;
- **Case 2**: a training subject performs actions not included in the training set (*Walking, Waving hand*) far from the camera (second picture in **Figure 5C**);
- **Case 3**: observing the shadow of an action included in the training set (*Pointing*) (third picture in **Figure 5C**) and a new one characterized by a whole-body motion (*Walking*) as opposed to the upper-body motions considered in the training set.

The accuracies are reported in the graph in **Figure 5C**. Cases I and III show how our method is tolerant to the presence of severe occlusions and, to some extent, is able to deal with indirect information, such as the one produced by the shadow of a moving object. As expected, both situations produce good results, with a relatively small decay in the performances. On the contrary, Case II shows a greater decrease in performance, probably due to the too limited extension of the apparent motion caused by the large distance of the motion from the camera.

4.2.4. Novel Dynamic Events

We consider here actions executed on the fronto-parallel plane and movements performed in depth, on a transverse plane. As for fronto-parallel dynamics, we focus in particular on handwriting, considering the following subcategories: frontal drawings of smooth symbols (as ellipses, infinite, see the first picture in **Figure 5D**, **Case 1**), hearts, (**Case 2**), sharp symbols (as rectangles and lines, **Case 3**), unconstrained text writings (**Case 4**).

Concerning the movements in depth, we identified the following scenarios: a user performing natural, unconstrained movements (**Case 5**, see an example in the second picture of **Figure 5D**); drawing smooth symbols on a table (**Case 6**, see an example in the third picture of **Figure 5D**); drawing hearts on a table (**Case 7**); drawing of sharp symbols on a table (**Case 8**); free text writing on a table (**Case 9**); and natural movements toward the robot (hi5, handshake, **Case 10**).

We considered both smooth and sharp shapes in order to test the method in case of continuous movements similar to the ones on which the Two-Thirds Power Law has been already tested in the literature (smooth shapes), and in case of other types of movements as the discontinuous ones (sharp shapes). The accuracies are reported in the graph of **Figure 5D**. We can observe a very good accuracy in the fronto-parallel cases (from Case 1 to Case 4). Regarding the movements in depth, we can observe that there is a decay in the performances in Case 5, as it includes very different movements with respect to training, with some even involving complex forces (like in the action of hammering); Case 6, the actions of drawing smooth shapes, shows a very good performance, while it decreases in Case 7 and Case 8, respectively, the actions of drawing hearts and sharp shapes; the accuracy is very good in Cases 9 and 10, respectively, the action of writing on a table, which indeed was in the training set (even if the video has been acquired in different place and time), and the actions toward the robot.

4.2.5. View-Point Changes with Different Sensors

We consider the same movements adopted for Test IV, but observed from an opposite point of view and using two different sensors (a common webcam and the camera Canon EOS 550D). The use of different sensors and the change of perspective lead to the generation of optical flow fields that may differ significantly from the ones adopted for the training phase.

We organized the tests considering the same classification adopted in Test IV. Planar movements have been observed with a webcam (320×240 pixels, 20 fps), while the sequences of actions in depth have been acquired with the Canon (320×240 pixels, 30 fps).

The accuracies are reported in the graph of **Figure 5E**. We can observe that changing the point of view there is a decay in the performance in the fronto-parallel cases (Case 1–Case 4), except for a small increase in Case 3, while there is an increase in the performance in all the cases of movements in depth except for Case 10. The movements with inflection points (drawing movements of smooth shapes, both on the fronto-parallel plane and in depth), which should be robust to the change in the point of view, are Case 1 (decrease of 0.29) and Case 6 (increase of 0.13).

5. IMPLEMENTATION IN THE iCub FRAMEWORK

Given the promising results derived from the offline testing reported in the previous sections, we propose a version of the method able to work online and to be integrated on the software framework of the robot iCub. The final goal is to embed the human activity detection system in a more structured architecture supporting natural attention redeployment and gaze behavior by the robot.

The software framework of the solution is designed to leverage the modularity supported by the middleware Yarp (Metta et al., 2010b) to enable two different computation stages: the

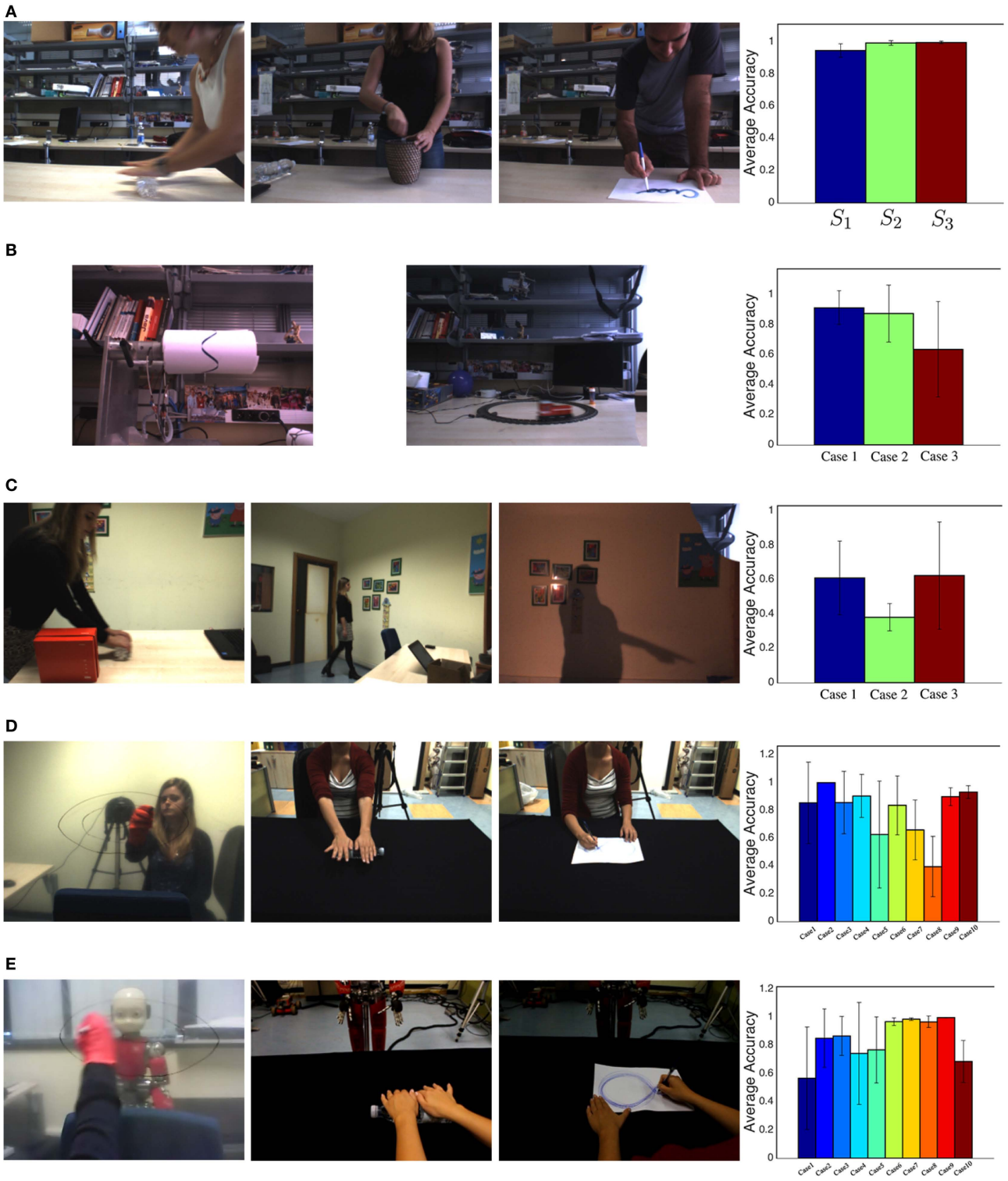

FIGURE 5 | Frames from sequences adopted in the Tests and graphs with the average classification accuracy (see the text for details on different cases for each Test). The images are the ones that have been really used for the analysis: the low quality is due to both the resolution and the bad lighting condition. In some cases, the presence of a transparent board placed between the camera and the observed scene further affect the quality of the images. **(A)** Test I on new videos of biological movements observed during training. Each bar refers to a training subject. **(B)** Test II on movements with changing speed. **(C)** Test III on videos with occlusions, far events and shadows. **(D)** Test IV on novel dynamic events. **(E)** Test V on novel dynamic events observed from a different view-point.

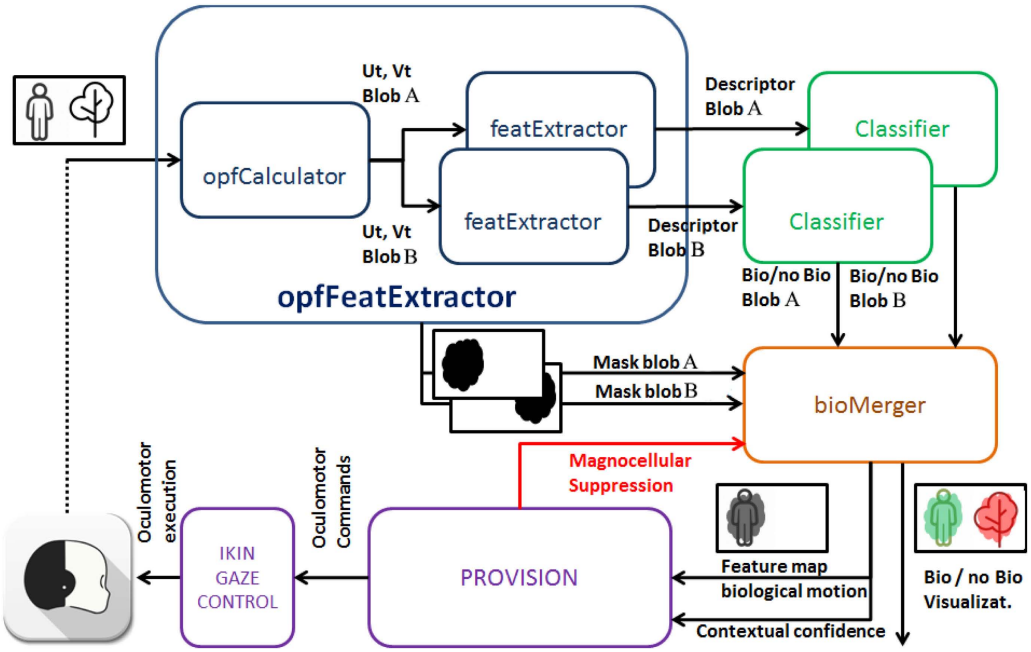

FIGURE 6 | The iCub architecture where our method has been implemented.

perception of biological movement and the synthesis of biological oculomotor actions. Modularity guarantees optimal computation distribution on the network resources and scalability of the solution. In **Figure 6**, we show the structure of the framework indicating how interconnections between modules closes the sensor-action loop through the execution of oculomotor actions based on salient loci in the stream of input images.

In the following, we review each module in detail. Although our solution may account for a generic number N of moving entities in the scene, without loosing in generality, we focus on the case $N = 2$ to exemplify the system behavior.

5.1. OpfFeatExtractor

The module resembles the early stage of visual pathways analyzing images of size 320×240 acquired from the eye cameras. The module comprises two classes of parallel computing, opfCalculator and featExtractor, which with reference to Section 3 correspond to the functionalities of motion segmentation and description, respectively. The parallelization of the necessary computation demand in multiple threads guarantees efficiency and real-time performance.

Two instances of the featExtractor class analyze the most salient and persistent blobs in the image plane, henceforth named A and B blob. The correctness of the data transferring from the opfCalculator to the featExtractor is guaranteed by supervised access (Yarp::Sig::Semaphore) to the two shared resources, $\text{srA} = [U_t, V_t, blobA]$ and $\text{srB} = [U_t, V_t, blobB]$. The opfCalculator module provides maps of the horizontal (U_t) and vertical (V_t) components of the optical flow on the whole image and the masks of A ($blobA$) and B ($blobB$) blobs to the rest of the network *via* tcp ports. In addition to the blob descriptors, the opfFeatExtractor also provides two monocromatic images, the binary maps marking the locations of blobs A and B in the image plane.

5.2. Classifier

The Classifier is a module that wraps few novel functions around the Machine Learning library GURLS (Tacchetti et al., 2013). The module is programmable from remote (RPC port) allowing the user to direct the modules, triggering different functionalities, the most relevant being the *training* of the model, and the *online recognition* to classify new observations. When model training functionality is activated, information coming from the opfFeatExtractor module is collected in a training set. When an appropriate amount of data is available, the module invokes a GURLS function to train a binary classifier using RLS (see Section 3.3). After the training, the model is adopted for online recognition, when at each time instant, new observed stimuli are described and classified. Classification is instantaneously based on the RLS score, generating a vote for the biological class if the score is positive, or for the non-biological class in case it is negative.

To partially correct instability of the final classification due to temporary failures, votes are collected into a temporal buffer of size 15. At each time instant, the final output of the classifier is based on a statistic of the votes in the buffer: when the majority of them (at least the 60%) is for a certain class, then the new observed event is labeled as an instance of that class, otherwise the system returns a temporary uncertain response as feedback.

5.3. BioMerger

The bioMerger module synchronizes the feedbacks of two classifiers (both A and B classifier) with the binary masks provided by the opfFeatExtractor module. Consequently, the module generates a color image of size 320×240 where the detected blobs A and B are color-coded according to their associated labels as depicted in **Figures 7B,D,F,H**. In addition, the module prepares a topographic feature map designed to compete in the visual attention system PROVISION (PROactive VISion attentiON)

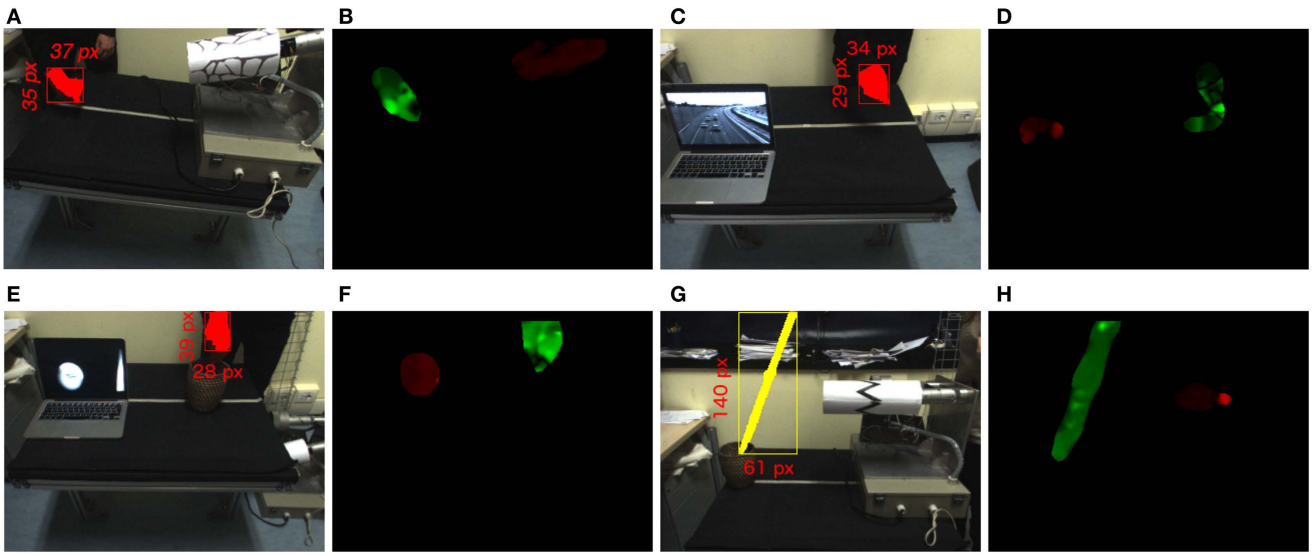

FIGURE 7 | (A,C,E,G) Representation of the setup and color segmentation for biological motion samples. **(B,D,F,H)** Biological (green) and non-biological (red) movements detected in panels **(A,C,E,G)**.

(Rea et al., 2014). The spatial map (320 × 240 grayscale image) indicates as saliency the spatial locations where the biological movement is detected. Such a rate is compliant with the temporal dynamics of saccades in human attentive systems, and the process as a whole resembles the infant predisposition to bias attention toward biological movement in the scene. The bioMerger streams a top-down command to the attentive system modifying the weight of biological motion in the competition for the attention, according to the confidence of the classification.

5.4. PROVISION

The PROVISION is log-polar attention system based on the computation model for attentive systems proposed in the study of Itti et al. (1998). Through the combination of two fundamental processes, Winner-Take-All (WTA) and Inhibition-of-Return (IoR), the visual attention selects the most significant location in the saliency map. The selection of the saliency winning location activates a ballistic oculomotor action (saccade) that brings the salient stimulus in the camera center of the drive eye. In this contribution, we enhanced the collection of feature maps with an additional feature map responding to the presence of biological motion in the image plane. The mechanism triggers PROVISION autonomous focus of attention redeployment toward the biological movement which in turn triggers an oculomotor command to the IkinGazeControl. The PROVISION system provides to the rest of the network a command of suppression of the movement perception. The process resembles the suppression of the magnocellular visual pathway (Burr et al., 1994) and guarantees to avoid excessive activation of the visual pathway caused by the egomotion during the saccade. The bioMerger leverages the PROVISION command of suppression to idle the process of extraction of the biological movement feature. This assures a stable perception–action loop comprising the extraction of the optical flow, the classification, and the execution of oculomotor actions, such as saccades.

5.5. IkinGazeControl

The biological control (Roncone et al., 2016) accounts for both the neck and eye control. The combination of two independent controls guarantees the convergence of the fixation point on the target. The controller solves the fixation tasks by implementing a biologically inspired kinematic controller that computes the robot joint velocities in order to generate minimum-jerk, quasi-straight trajectory of the fixation point. The controller is also enriched with additional models of biological oculomotor actions such as vestibular ocular reflex and passive gaze stabilization. The PROVISION system gives instructions to the IkinGazeController that autonomously coordinates 3 degrees-of-freedom (DoF) neck and 3-DoF eye system to show natural behavior in the robot gaze.

6. THE METHOD AT WORK ON THE ROBOT

In this section, we present the experimental analysis performed online on the robot. We start analyzing the accuracy for the classification of biological motion, even in the presence of different moving stimuli in the observed scene. Later, we will discuss the integration with the attention system and biological control system of the oculomotor action in the humanoid robot iCub.

6.1. Experiments on Online Learning

We describe the classification performances obtained on the robot. We first observe that, in typical applications involving proactive robots, it is fundamental to provide reliable training and classification in a reasonable time span. In the reported experiment, we show how this is achieved by parallelizing tasks in the software infrastructure.

To facilitate reproducibility both the biological and the non-biological stimuli are presented on a table (64 cm of height) and at a distance of 64 cm from the origin of the iCub frame of reference.

Training is performed starting from an initial condition without *a priori* knowledge, meaning that the robot lacks the abilities of discriminating between biological and non-biological motion. The training is performed online on the robot, replicating the situation where the operator interactively instructs the robot. Model selection is also performed online.

We first train the robot on the set of biological and non-biological categories already adopted for the offline analysis (see Section 4) using the multi-resolution temporal scheme selected in that circumstance. On average, each video lasts about $20''$. We test the classification system by proposing a single stimulus from a subset of representative event categories in different portions of the iCub field of view. The obtained results are shown in the first part of **Table 3**. During the evaluation, we determine whether each received packet matches the expected response (column *Accuracy A*). The reported accuracy is obtained in asynchronous evaluation periods (column *Time*). A relevant aspect for robotic applications, requiring adaptability to context change in the environment, is the transmission rate (column *Rate A*), which is reasonable for oculomotor actions such as saccades.

Finally, we consider an experimental scenario where two stimuli (A and B) are presented on different portions of the field of view. The analysis of the classification quality is reported in the second part of **Table 3**, where we show accuracy and transmission rate corresponding to the A and B stimulus, plus the overall evaluation time. The classification of the motion is uncertain at the initial transient, due to the filtering of the features vector and the instability of the classification. In the table, we report the accuracy excluded the initial time windows necessary for stable filtering of the result.

Despite that the number of classifiers has increased, the decrease in the rate is limited and it has no effect on the gaze control. In fact, the software framework is designed to be scalable and the computation demand of multiple classifiers is distributed across the processing node in the network. Overall, this set of experiments produces convincing results for what concerns accuracy. We only have a degradation on the pair *leaves* + *rolling dough* stimuli, due to discontinuities of the stimulus provided.

Figure 8A shows how the classifier generates response messages, for a biological stimulus (*rolling dough*). The score provided by the classifier is accumulated over a 15-frame temporal window. In this case, the response is constantly 1.0 indicating a correct classification as biological movement. The brief undetermined classification (classifier response: 0.0) is due to scores below zero in the previous window of 15 frames, as depicted in picture. The system recovers after few iterations and the classification returns to provide correct response giving evidence of robustness.

6.2. Experiment on Integration with PROVISION and Gaze Control

The classification system is designed to reliably provide results to a broad range of software applications in the iCub network. To facilitate its use, we integrate the classification output with the masks produced by the opfFeatExtractor into a single mask. The mask is produced and provided to the network by our

TABLE 3 | Online classification results with one stimulus and two stimuli.

Stimuli	Accuracy A	Rate A (pkt/s)	Accuracy B	Rate B (pkt/s)	Time (s)
Clouds	1.0	4.30	–	–	40
Leaves	0.94	4.40	–	–	40
Rolling dough	0.96	4.02	–	–	40
Transporting	1.0	4.19	–	–	40
Clouds-rolling dough	1.0	3.13	1.0	3.13	30
Clouds-transporting	0.98	3.27	1.0	3.26	30
Leaves-rolling dough	0.81	3.51	0.90	3.57	30
Leaves-transporting	0.85	3.83	1.0	4.11	30

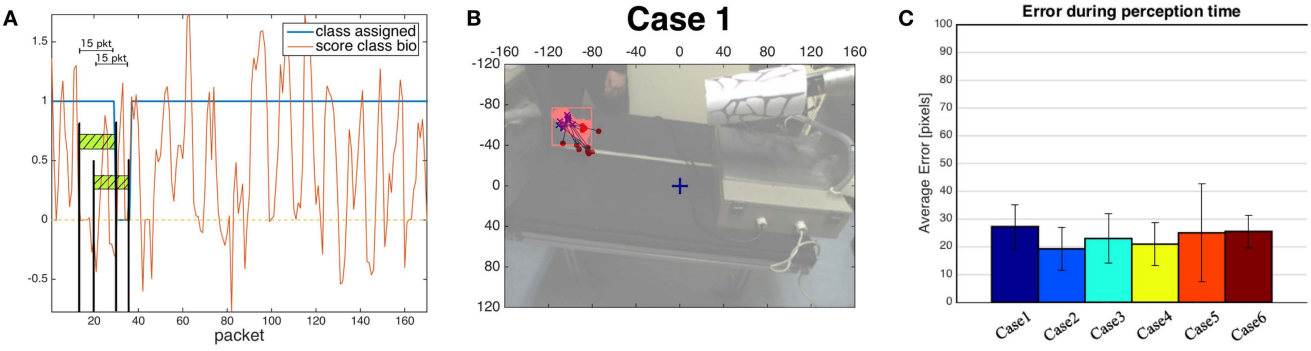

FIGURE 8 | (A) An example of the classifier generating response messages—*Rolling dough* case (bio = 1, non-bio = −1, "?" = 0). **(B)** An example of perception errors, computed as the distances between the locations identified by the color segmentations (blue crosses) and the corresponding positions individuated by the biological motion detector (red dots). The dots and the crosses are printed on a more transparent version of the scene extracted from the robot point of view before the beginning of the oculomotor action. **(C)** Average perception error in all cases. For the detailed list see **Table 4**.

new bioMerger module. In this experiment, the output of the bioMerger module interfaces with pre-existing software: PROVISION and iKin Gaze Control (Pattacini et al., 2010). The biological movement detector provides a feature map of biological movement and the level of confidence associated with classification.

The integration experiment described here includes two different stages: perception and action. For both the evaluation stages, we produce a biological and a non-biological movement (distractor) and we determine how the position of the salient biological stimulus evolves over time. To determine the ground truth for the localization of the human activity in the scene, we adopt a color segmentation module and we perform experiments with a human subject wearing colored gloves. The localization based on color relies on a source of information alternative to the one exploited by our algorithm (i.e., motion), thus representing a dependable estimation for comparison (see **Figure 7**). For each case, beyond measuring the perceptual error, we run multiple saccades and extract the statistics on the errors due to the control stage.

In the evaluation of perception quality, we compare our estimated (u, v) position of the salient stimulus in the image plane provided by PROVISION with the segmentation of the moving region detected by the color segmentation (**Figure 8B**). In this phase, no oculomotor command is generated, and the fixation point of the robot is at the center of the scene in $F = [-0.5, 0.0, -0.35]$m where the frame of reference is located and oriented according to the iCub standards. In **Figure 8C**, we show the distance in pixels between the two different locations. Notice that a mean distance in the range [20–40] pixels corresponds to a metric range [4–8] cm, given the distance of the camera from the stimuli (64 cm).

In the evaluation of the action quality, we analyze how the biological movement detector biases the proactive attentive system of the humanoid robot iCub. The visual attentive system generates a saccade command and once the controller plans the relative saccade, the oculomotor action is executed bringing the center of the robot eye (fovea) to the most salient stimulus, winning in the competition between perceptual features. Considering the known distance of the stimulus from the stereo cameras, the gaze controller moves the fixation point to the target of interest (the biological movement in the scene). In PROVISION, a postsaccadic refinement mechanism based on visual feedback control can potentially refine the saccade. However, we disabled such additional control to avoid unclear measurements on two distinct and concurrent visual processes on the robot. Notice that, as shown in **Table 4**, the system performs incorrect saccades in two different cases. The two incorrect saccades are due to a misclassification of the biological stimulus by the Classifier module: they have been discarded for both the evaluation of the perception and action quality.

In **Figures 9A–F**, we show the position given by the color segmentation when the saccade starts (blue crosses), and the trajectory of the position given by the saccadic commands (red line) toward the fovea (0,0), from when the saccade starts up to 2.5 s (as, from **Figures 9G–L**, where we represent the velocity of the fixation point during the executing of the saccade, in other words during its approaching to the target, we can consider this time as the duration of the saccade). The semitransparent image overlapping the trajectory is to consider as the snapshot of one visual scene taken right before the oculomotor command saccade is triggered.

TABLE 4 | Number of correct saccades in integration experiment.

Case	Stimuli (A, B)	Corr/tot
Case 1	**Gesturing**-wheel random	11/11 sac.
Case 2	Leaves-**writing subject 1**	10/11 sac.
Case 3	Cars-**gesturing**	15/15 sac.
Case 4	Bouncing ball-**mixing**	11/12 sac.
Case 5	**Mixing, no person**-wheel zigzag	11/11 sac.
Case 6	Wheel random-**writing subject 2**	13/13 sac.

Bold font refers to the biological actions.

We measure the control error by computing the distance between the center of the fovea (0,0) and the position given by the saccadic command (red line of **Figures 9A–F**) for the six typologies of trials in the previous perception stage. In **Figures 10A–F**, we show the error from the moment the saccade starts up to about 6 s. In the graphs of the control error the mean of the error (blue solid line) reaches immediately the quality threshold of 40 pixels. The threshold is set according to our estimation that at a distance of 0.68 m, 40-pixel error is interpreted as a correct saccade from a human observer. The responses in Case 1, Case 3, and Case 5 show overshoots, due to the relative position of the biological stimulus with respect to the resting position. All the responses converge to control errors below the 40-pixel threshold guaranteeing the expected quality of the control of the saccade to the biological movement. The oscillations after the transient are due to the response of the color segmentation that tracks moving stimuli after the end of the saccade and it is not related to the quality of the saccade generated by the system.

Then, we measure the distance between the center of the fovea (0,0) and the centroid of the color detection system (**Figures 11A–F**): this can be referred to as a global error, as it includes both the perception error and the control error.

Case 5 is a very peculiar case as the color segmentation gives us the position of the center of the rectangle around the stick (**Figure 9E**), while our module will give as oculomotor command the most salient position of the saliency map, which corresponds to the position of maximum optical flow. This leads to a larger SD in the perception error (**Figure 8C**, Case 5) and a larger control error (**Figure 10E**). However, considering the goal of detecting humans in the scene, our method could be considered actually more accurate than color segmentation.

7. FINAL DISCUSSION

In this paper, we presented a computational model for discriminating between biological and non-biological movements in video sequences, leveraging a well-known regularity of human motor control. Notwithstanding the large heterogeneity of the dynamics of the motions that can be encountered in everyday life situations, we proposed a temporal multi-resolution descriptor, purely based on low-level motion features. We showed that this descriptor has on average a better performance than any single-resolution

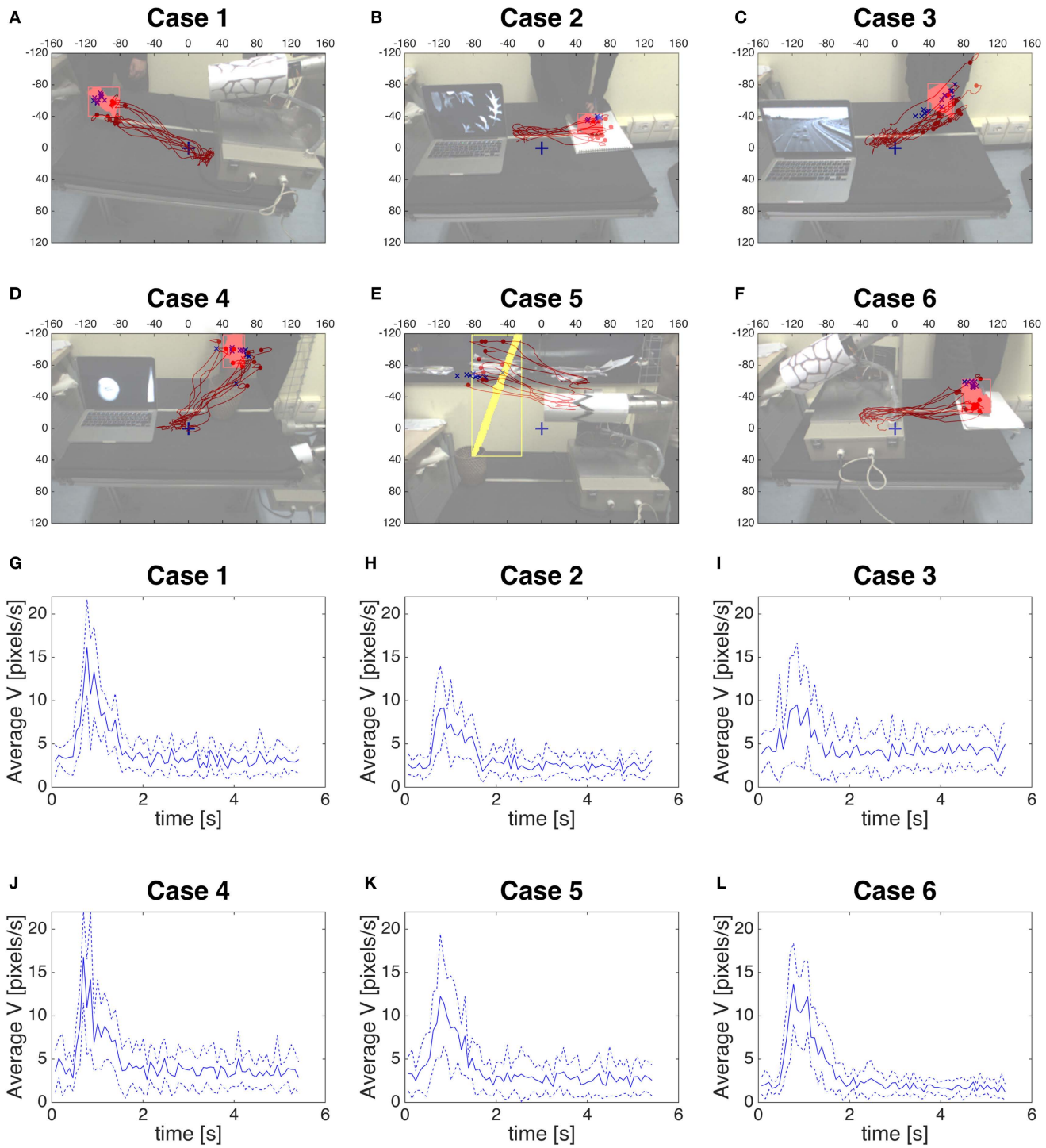

FIGURE 9 | (A–F) The position given by the color segmentation when the saccade starts (blue crosses), and the trajectory given by the saccadic commands (red line) toward the fovea (0,0). **(G–L)** The blue line is the average velocity of the fixation point to reach the target for each case, the blue dotted line is the average velocity ± SD.

descriptor, as the latter fails in capturing the large variability of possible dynamics of the motions.

We demonstrated the descriptor to be effective also for events of a variable temporal duration and to generalize well to new and challenging scenarios. It should be noticed that our approach does not require any appearance-based detection of the human partner, as the regularities of biological motion are extracted independently of the agent's shape. This feature guarantees the possibility to recognize human activities also when the agent is not visible or severely occluded, e.g., observing a shadow or a

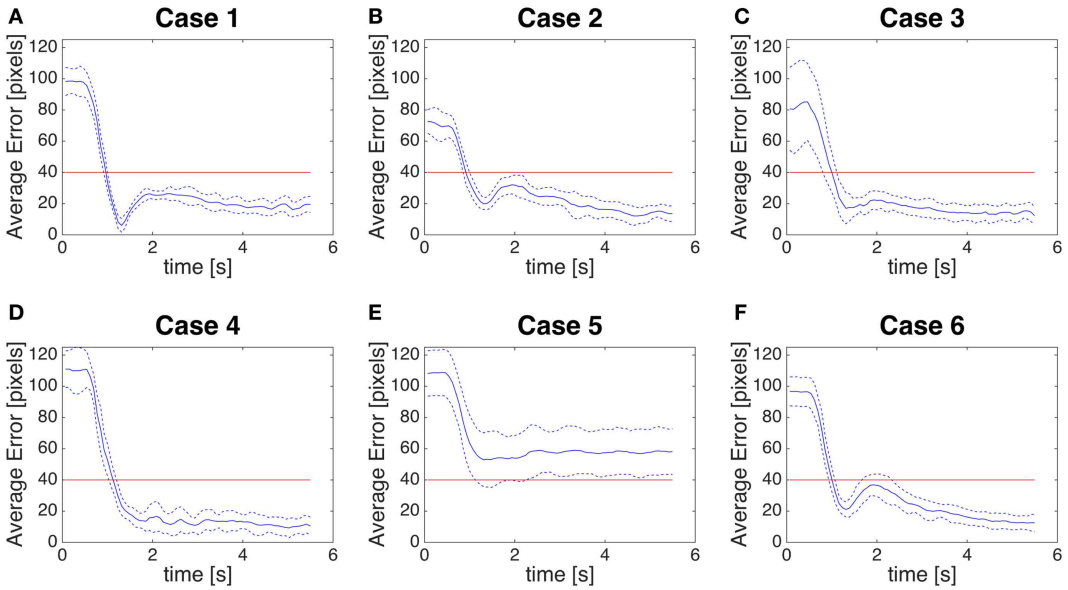

FIGURE 10 | (A–F) Control error. Distance between the position given by our module as saccadic command and the fovea (0,0). The blue line is the average error, the blue dotted line is the average error ± SD. The red line is a threshold of 40 pixels.

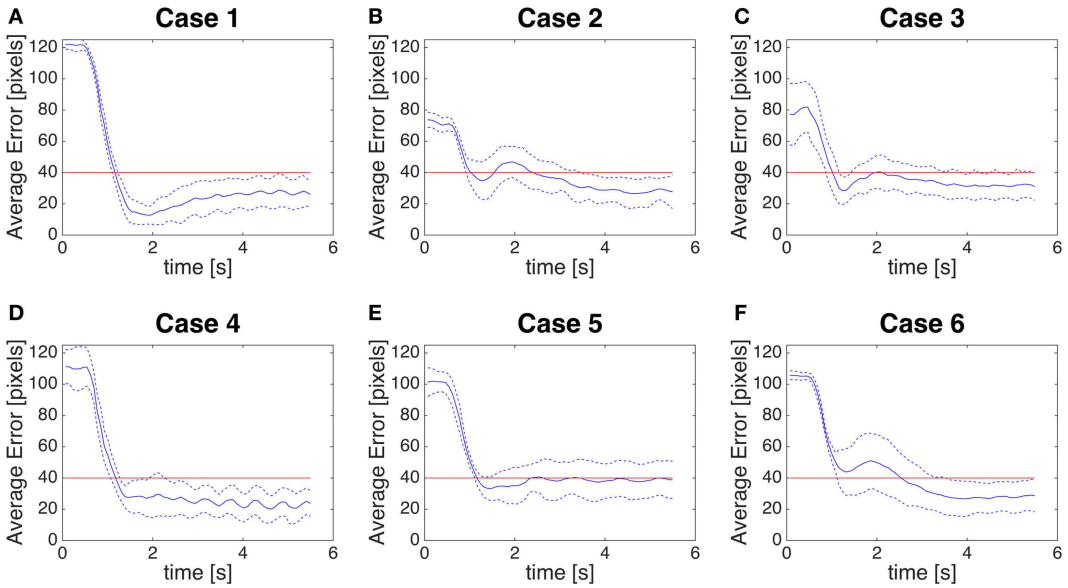

FIGURE 11 | (A–F) Global error. Distance between the centroid of the biological motion (given by color segmentation) and the fovea (0,0). The blue line is the average error, the blue dotted line is the average error ± SD. The red line is a threshold of 40 pixels.

visible tool moved by a hidden agent. The proposed model could be exploited in industrial applications, for instance, in assembly lines tasks, with the aim of distinguishing between the movement of a human operator and an object moved by the conveyor belt. Another possible setting of application is for traffic control, to distinguish motion caused by pedestrians and cars for monitoring purposes.

This computational model can, therefore, enable an artificial agent to detect the presence of humans in its surrounding to provide the appropriate pro-social behavior, as we demonstrate by implementing it on the humanoid robot iCub. The video at this link shows some real-world experiments of the proposed computational model.

The saccadic action performed by the robot as a consequence of the detection of human activity in the scene, beyond providing the robot with a better view on the area where it is more probable that an interaction could start, also informs the human partner in an intuitive way about the internal attentional status of the robot.

This type of gaze-based intuitive communication, commonly adopted in conversational agents and social robotics, has recently gained impact also in the field of small manufacturing, where Baxter (by Rethink robotics) exploits a screen with (non-functional) eyes, just to reveal its focus of attention. In our system, the matching between the actual function of the eyes (i.e., cameras) and their ostensive value increases even more the intuitive interpretation of the iCub actions.

In this respect, our work represents the first building block of the social abilities of the robot, which in the future will be exploited to categorize actions into different classes, an issue that we have started to address in Noceti et al. (2015a). Such capability can be of strategic interest for a broad community aiming at enabling effective interaction between human, robots and intelligent machines.

I, Alessia Vignolo hereby declare that all the photos in the manuscript are of myself and of collaborators, and we all consent for the photos to be published in the manuscript.

AUTHOR CONTRIBUTIONS

All the authors contributed to the design of the experiment, AV cured the data analysis and collection, and all the authors contributed to the writing of the manuscript.

FUNDING

This work was supported by the European CODEFROR project (FP7-PIRSES-2013-612555).

REFERENCES

Asendorpf, J., and Baudonniere, P. (1993). Self-awareness and other-awareness: mirror self-recognition and synchronic imitation among unfamiliar peers. *Dev. Psychol.* 29, 88–95. doi:10.1037/0012-1649.29.1.88

Bisio, A., Sciutti, A., Nori, F., Metta, G., Fadiga, L., Sandini, G., et al. (2014). Motor contagion during human-human and human-robot interaction. *PLoS ONE* 9:e106172. doi:10.1371/journal.pone.0106172

Boucher, J. D., Pattacini, U., Lelong, A., Bailly, G., Elisei, F., Fagel, S., et al. (2012). I reach faster when I see you look: gaze effects in human-human and human-robot face-to-face cooperation. *Front. Neurorobot.* 6:3. doi:10.3389/fnbot.2012.00003

Breazeal, C., Edsinger, A., Fitzpatrick, P., and Scassellati, B. (2001). Active vision for sociable robots. *IEEE Trans. Syst. Man Cybern. A Syst. Hum.* 31, 443–453. doi:10.1109/3468.952718

Breazeal, C., and Scassellati, B. (1999). "A context-dependent attention system for a social robot," in *IJCAI International Joint Conference on Artificial Intelligence* (San Francisco, CA: Morgan Kaufmann Publishers Inc), 1146–1151.

Brethes, L., Menezes, P., Lerasle, F., and Hayet, J. (2004). "Face tracking and hand gesture recognition for human-robot interaction," in *Robotics and Automation, 2004. Proceedings. ICRA '04. 2004 IEEE International Conference on*, Vol. 2, New Orleans, LA, 1901–1906.

Burr, D., Morrone, M., and Ross, J. (1994). Selective suppression of the magnocellular visual pathway during saccadic eye movements. *Nature* 371, 511–513. doi:10.1038/371511a0

Chaminade, T., and Cheng, G. (2009). Social cognitive neuroscience and humanoid robotics. *J. Physiol. Paris* 103, 286–295. doi:10.1016/j.jphysparis.2009.08.011

Correa, M., Hermosilla, G., Verschae, R., and Ruiz-del Solar, J. (2012). Human detection and identification by robots using thermal and visual information in domestic environments. *J. Intell. Robot. Syst.* 66, 223–243. doi:10.1007/s10846-011-9612-2

Dillmann, R. (2004). Teaching and learning of robot tasks via observation of human performance. *Rob. Auton. Syst.* 47, 109–116. doi:10.1016/j.robot.2004.03.005

Elsner, C., Falck-Ytter, T., and Gredebäck, G. (2012). Humans anticipate the goal of other people's point-light actions. *Front. Psychol.* 3. doi:10.3389/fpsyg.2012.00120

Farnebäck, G. (2003). "Two-frame motion estimation based on polynomial expansion," in *Image Analysis. SCIA 2003. Lecture Notes in Computer Science*, eds J. Bigun and T. Gustavsson (Berlin, Heidelberg: Springer), 2749.

Farroni, T., Csibra, G., Simion, F., and Johnson, M. H. (2002). Eye contact detection in humans from birth. *Proc. Natl. Acad. Sci. U.S.A.* 99, 9602–9605. doi:10.1073/pnas.152159999

Gaschler, A., Jentzsch, S., Giuliani, M., Huth, K., de Ruiter, J., and Knoll, A. (2012). "Social behavior recognition using body posture and head pose for human-robot interaction," in *IEEE/RSJ International Conference on Intelligent Robots and Systems*, Vilamoura, 2128–2133.

Gavazzi, G., Bisio, A., and Pozzo, T. (2013). Time perception of visual motion is tuned by the motor representation of human actions. *Sci. Rep.* 3, 1168. doi:10.1038/srep01168

Greene, P. H. (1972). Problems of organization of motor systems. *Prog. Theor. Biol.* 2, 123–145.

Itti, L., Koch, C., and Niebur, E. (1998). A model of saliency-based visual attention for rapid scene analysis. *IEEE Trans. Pattern Anal. Mach. Intell.* 11, 1254–1259. doi:10.1109/34.730558

Koppula, H. S., Gupta, R., and Saxena, A. (2013). Learning human activities and object affordances from rgb-d videos. *Int. J. Robot. Res.* 32, 951–970. doi:10.1177/0278364913478446

Lacquaniti, F., and Terzuolo, C. (1983). The law relating the kinematic and figural aspects of drawing movements. *Acta Psychol.* 54, 115–130. doi:10.1016/0001-6918(83)90027-6

Lacquaniti, F., Terzuolo, C., and Viviani, P. (1983). The law relating the kinematic and figural aspects of drawing movements. *Acta Psychol.* 54, 115–130. doi:10.1016/0001-6918(83)90027-6

Méary, D., Kitromilides, E., Mazens, K., Graff, C., and Gentaz, E. (2007). Four-day-old human neonates look longer at non-biological motions of a single point-of-light. *PLoS ONE* 2:e186. doi:10.1371/journal.pone.0000186

Metta, G., Natale, L., Nori, F., Sandini, G., Vernon, D., Fadiga, L., et al. (2010a). The icub humanoid robot: an open-systems platform for research in cognitive development. *Neural Netw.* 23, 1125–1134. doi:10.1016/j.neunet.2010.08.010

Metta, G., Natale, L., Nori, F., Sandini, G., Vernon, D., Fadiga, L., et al. (2010b). The icub humanoid robot: an open-systems platform for research in cognitive development. *Neural Netw.* 23, 1125–1134. doi:10.1016/j.neunet.2010.08.010

Mori, M. (1970). The uncanny valley. *Energy* 7, 33–35.

Mutlu, B., Kanda, T., Forlizzi, J., Hodgins, J., and Ishiguro, H. (2012). Conversational gaze mechanisms for humanlike robots. *ACM Trans. Interact. Intell. Syst.* 1, 12:1–12:33. doi:10.1145/2070719.2070725

Noceti, N., Sciutti, A., Rea, F., Odone, F., and Sandini, G. (2015a). "Estimating human actions affinities across views," in *Proceedings of the 10th International Conference on Computer Vision Theory and Applications (VISIGRAPP 2015)*, Berlin, Germany, 130–137.

Noceti, N., Sciutti, A., and Sandini, G. (2015b). "Cognition helps vision: recognizing biological motion using invariant dynamic cues," in *Image Analysis and Processing – ICIAP 2015. ICIAP 2015. Lecture Notes in Computer Science*, eds V. Murino and E. Puppo (Cham: Springer), 9280.

Palinko, O., Sciutti, A., Schillingmann, L., Rea, F., Nagai, Y., and Sandini, G. (2015). "Gaze contingency in turn-taking for human robot interaction: advantages and drawbacks," in *24th IEEE International Symposium on Robot and Human Interactive Communication* (Kobe, Japan).

Pattacini, U., Nori, F., Natale, L., Metta, G., and Sandini, G. (2010). "An experimental evaluation of a novel minimum-jerk cartesian controller for humanoid robots," in *IEEE/RSJ International Conference on Intelligent Robots and Systems* (Taipei: IEEE), 1668–1674.

Pozzo, T., Papaxanthis, C., Petit, J. L., Schweighofer, N., and Stucchi, N. (2006). Kinematic features of movement tunes perception and action coupling. *Behav. Brain Res.* 169, 75–82. doi:10.1016/j.bbr.2005.12.005

Rea, F., Muratore, P., and Sciutti, A. (2016). "A 13-year-olds approach human-robot interaction like adults," in *IEEE International Conference Developmental Learning and Epigenetic Robotics*, Cergy-Pontoise.

Rea, F., Sandini, G., and Metta, G. (2014). "Motor biases in visual attention for a humanoid robot," in *IEEE/RAS International Conference of Humanoids Robotics*, Madrid.

Richardson, M., and Flash, T. (2002). Comparing smooth arm movements with the two-thirds power law and the related segmented-control hypothesis. *J. Neurosci.* 22, 8201–8211.

Roncone, A., Pattacini, U., Metta, G., and Natale, L. (2016). "A Cartesian 6-dof gaze controller for humanoid robots," in *Proceedings of Robotics: Science and Systems* (Ann Arbor, MI).

Sciutti, A., Bisio, A., Nori, F., Metta, G., Fadiga, L., Pozzo, T., et al. (2012). Measuring human-robot interaction through motor resonance. *Int. J. Soc. Robot.* 4, 223–234. doi:10.1007/s12369-012-0143-1

Simion, F., Regolin, L., and Bulf, H. (2008). A predisposition for biological motion in the newborn baby. *Proc. Natl. Acad. Sci. U.S.A.* 105, 809–813. doi:10.1073/pnas.0707021105

Sung, J., Ponce, C., Selman, B., and Saxena, A. (2012). "Unstructured human activity detection from rgbd images," in *IEEE International Conference on Robotics and Automation* (Saint Paul, MN), 842–849.

Tacchetti, A., Mallapragada, P. K., Santoro, M., and Rosasco, L. (2013). Gurls: a least squares library for supervised learning. *J. Mach. Learn. Res.* 14, 3201–3205.

Tomasello, M., Carpenter, M., Call, J., Behne, T., and Moll, H. (2005). Understanding and sharing intentions: the origins of cultural cognition. *Behav. Brain Sci.* 28, 675–691. doi:10.1017/S0140525X05000129

Urgen, B. A., Plank, M., Ishiguro, H., Poizner, H., and Saygin, A. P. (2012). "Temporal dynamics of action perception: the role of biological appearance and motion kinematics," in *34th Annual Conference of the Cognitive Science Society*, Sapporo, 2469–2474.

Vieilledent, S., Kerlirzin, Y., Dalbera, S., and Berthoz, A. (2001). Relationship between velocity and curvature of a human locomotor trajectory. *Neurosci. Lett.* 305, 65–69. doi:10.1016/S0304-3940(01)01798-0

Vignolo, A., Noceti, N., Sciutti, A., Rea, F., Odone, F., and Sandini, G. (2016a). "The complexity of biological motion. A temporal multi-resolution motion descriptor for human detection in videos," in *IEEE International Conference Developmental Learning and Epigenetic Robotics*, Cergy-Pontoise.

Vignolo, A., Rea, F., Noceti, N., Sciutti, A., Odone, F., and Sandini, G. (2016b). "Biological movement detector enhances the attentive skills of humanoid robot iCub," in *IEEE-RAS International Conference on Humanoid Robots*, Cancun.

Viviani, P., Baud-Bovy, G., and Redolfi, M. (1997). Perceiving and tracking kinesthetic stimuli: further evidence of motor-perceptual interactions. *J. Exp. Psychol. Hum. Percept. Perform.* 23, 1232–1252. doi:10.1037/0096-1523.23.4.1232

Viviani, P., and Stucchi, N. (1992). Biological movements look uniform: evidence of motor-perceptual interactions. *J. Exp. Psychol. Hum. Percept. Perform.* 18, 603–623. doi:10.1037/0096-1523.18.3.603

Viviani, P., and Terzuolo, C. (1982). Trajectory determines movement dynamics. *Neuroscience* 7, 431–437. doi:10.1016/0306-4522(82)90277-9

Weinland, D., Ronfard, R., and Boyer, E. (2011). A survey of vision-based methods for action representation, segmentation and recognition. *Comput. Vis. Image Underst.* 115, 224–241. doi:10.1016/j.cviu.2010.10.002

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.