The AAAI 2017 Spring Symposium on
Science of Intelligence: Computational Principles of Natural and Artificial Intelligence
Technical Report SS-17-07

Computational Vision for Social Intelligence

Alessia Vignolo,[1,2] **Alessandra Sciutti,**[1] **Francesco Rea,**[1]
Nicoletta Noceti,[2] **Francesca Odone,**[2] **Giulio Sandini**[1]
[1]RBCS - Istituto Italiano di Tecnologia, Via Enrico Melen 83, 16152, Genova-IT
[2]DIBRIS - Universitá di Genova, via Dodecaneso 35, 16146, Genova-IT

Abstract

A fundamental trait of human intelligence is represented by social intelligence, which enables natural and fruitful interaction since very early during infancy. The ability to collaborate is also a key challenge for today's robotics, which could benefit from the design of computational models supporting the understanding of social intelligence for the future of human-robot interaction. Our research focuses on these topics from the perspective of computational vision. In particular we aim at understanding how social intelligence develops in presence of the very limited sensory-motor skills and prior knowledge common to babies. As a starting point we consider the natural predisposition of newborns to notice potential interacting partners in their surroundings, which is manifested by a preference for biological motion over other types of motion. To model this skill, we propose a video-based computational method for biological motion detection inspired by the *Two-Thirds Power Law*, a well-known invariant of human movements. In particular, we address the problem by recruiting machine learning framework, leveraging a binary classification to discriminate biological from non-biological stimuli from rather coarse motion models extracted from video measurements. After evaluating the performance of the method and its generalization power to complex scenarios in an offline test, the method is engineered to work online on a robot, the humanoid iCub. The integration with the attentional module of the robot enables it to direct its gaze toward human activity in the scene. We posit that the possibility for a robotic system to orient the attention toward potential interacting agents, as a human infant would, represents one of the first stages of social intelligence, on top of which more complex skills, as action and intention understanding, could emerge.

Introduction

A key challenge in current robotics has become to provide robots with social intelligence, to enable them to adapt to the complexity of real-world human interactions. 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 proficiently coordinate with their peers and caregivers, even in absence of language. Moreover, from the constrained social abilities exhibited in the very first months of life, humans are able to develop a full fledged social competence in adulthood. The essential social skills exhibited by newborns can therefore represent the minimum set of skills necessary to bootstrap more complex interactive abilities. In our line of research, we focus in particular on the natural predisposition of newborns to notice potential interacting partners in their surroundings, which is manifested by a preference for biological motion over other types of motion (Simion, Regolin, and Bulf 2008).

We propose a video-based computational method for biological motion detection by recruiting machine learning framework, leveraging a binary classification to discriminate biological from non-biological stimuli from rather coarse motion models extracted from video measurements. Although the question is clearly defined in machine learning terms, the heterogeneity of the data and the wide intra-class variability exacerbate the complexity of the task. In fact, the examples of biological motion we face in everyday life are characterized by very heterogeneous dynamics and trajectories, while the class of non-biological events is even less constrained, because it includes all motions not produced by a living being, such as vehicles, toys, or even natural elements affected by non-biological forces (e.g., the motion of the leaves caused by the wind).

To design a system sensitive to the regularities typical of biological movements we draw inspiration from the *Two-Thirds Power Law*, a well-known invariant of human movements describing the relationship between the instantaneous tangential velocity and the radius of curvature of human end-point movements (Lacquaniti and Terzuolo 1983). We choose this law according to the evidence that humans are sensitive to it since their first days after birth (Méary et al. 2007).

After evaluating the performance of our proposed method in correctly classifying biological and non biological dynamics from videos in an offline fashion, the possibility to exploit it for the online perception in intelligent systems is demonstrated by its implementation on the humanoid robot iCub (Metta et al. 2010), to guide robot attention toward potential interacting partners in the scene. The novel module extracts relevant features of biological motion with a computationally efficient algorithm and enriches the feature maps of a visual attentive system. The advantage of the solution is that robot attention is immediately biased towards human

activity in the scene even when the human agent is not directly visible.

Overall this process can be seen as a first step of a more complex behavioral architecture, that provides intelligent systems with deeper understanding of the observed action and effective planning of an interaction strategy towards human partners.

Methods

Offline Analysis

Our approach to discriminate between biological and non-biological dynamic events consists of three steps aiming at the *detection*, *representation* and *classification* of the biological motion in an observed scene.

Motion segmentation We start with a low-level analysis of the video stream (originated from one camera of the robot) to detect the moving regions in the scene. To this purpose, at each time instant, we first compute the dense optical flow map of an image (Farnebäck 2003). The optical flow map is thresholded to highlight only locations with a significant motion. To discard sporadic, noisy pixels responses, we apply a *perceptual grouping* operator, in which only locations whose neighboring pixels are also marked as moving are kept in the analysis. We obtain a *saliency map*, where we finally detect the connected components – the candidate regions for motion recognition.

Motion description At time t, let $(u_i(t), v_i(t))$ be the optical flow components associated with a point $\mathbf{p}_i(t)$ lying in a region $\mathcal{R}(t)$, and N the size of the region, i.e. the number of pixels within it. We compute a set of motion features which empirically estimate the analytical quantities related by the Two-Thirds Power Law (e.g., Tangential velocity: $\hat{\mathbf{V}}_i(t) = (u_i(t), v_i(t), \Delta_t)$, being Δ_t the temporal displacement between observations of two adjacent time instants. For a full list of the spatio-temporal dynamic features extracted see (Vignolo et al. 2016a)). The region $\mathcal{R}(t)$ is globally described with a feature vector $\mathbf{x}_t \in R^4$ by averaging the region contributions: $\mathbf{x}_t = \frac{1}{N}\left[\sum_i \hat{V}_i(t), \sum_i \hat{C}_i(t), \sum_i \hat{R}_i(t), \sum_i \hat{A}_i(t)\right]$. By correlating the regions over time, we may set up a temporal sequence of M observations $S_t \in R^{4M}$, $S_t = [\mathbf{x}_{t-M}, \ldots, \mathbf{x}_t]$, and apply a filtering (with a Gaussian mask in our case) to partially correct instantaneous noisy information that might affect the overall analysis. As a results we have at time t a filtered feature vector computed as $\tilde{\mathbf{x}}_t = S_t * G(M)$, where $*$ denotes the convolution, applied to each feature separately, with a Gaussian mask G of width equal to M. A more detailed description of the method and of its multi-resolution instantiation, which efficiently combines measurements that may span different temporal portions of an image sequence, can be found in (Vignolo et al. 2016a).

Motion classification We formulate the problem of recognizing biological motion from video sequences as a binary classification problem, which we address with a classical *Regularized Least Squares algorithm* (henceforth RLS). More in detail, we are given a training set $Z = \{(\tilde{\mathbf{x}}_i, y_i) \in X \times Y\}_{i=1}^n$, where $\tilde{\mathbf{x}}_i \in X = R^4$, while $y_i \in Y = \{-1, 1\}$ (label 1 refers to biological samples, label -1 indicates instead the non-biological samples). RLS amounts to minimize the following functional

$$f_Z = \min_{f \in \mathcal{H}} \frac{1}{n} \sum_{i=1}^{n} (y_i - f(\tilde{\mathbf{x}}_i))^2 + \lambda ||f||_{\mathcal{H}} \qquad (1)$$

where $\mathcal{H}$ is a Reproducing Kernel Hilbert Space for which the *representer theorem* holds: $f_Z^\lambda(\mathbf{x}) = \sum_{i=1}^n \alpha_i K(\mathbf{x}, \tilde{\mathbf{x}}_i)$, where $\alpha = (\mathbf{K} + n\lambda I)^{-1}\mathbf{y}$, with K a Mercer kernel and $\mathbf{K}$ the kernel matrix computed on the training set. In this work we consider a Radial Basis Function (RBF) kernel.

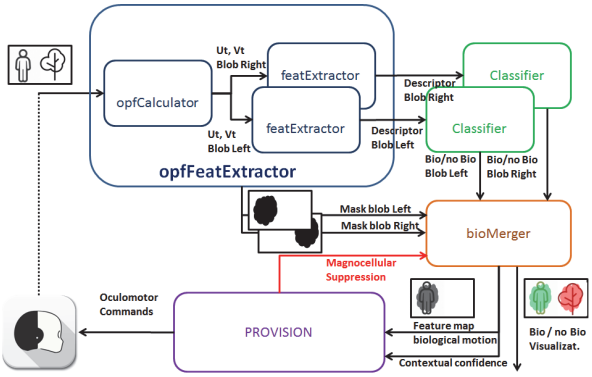

Figure 1: The iCub architectural framework implementing the proposed solution. From (Vignolo et al. 2016b)

Implementation in the iCub framework

The software framework adopted in the proposed solution is shown in figure 1. The **OpfFeatExtractor** resembles the early stage of visual pathways associated with the extraction of motion, processing the images acquired with the iCub's camera. It analyzes the most salient and persistent blobs on two half portions of the image plane and provides maps of the horizontal (U_t) and vertical (V_t) components of the optical flow [1]. The **Classifier** module is a wrapper around the Machine Learning library GURLS (Tacchetti et al. 2013) supporting the *training* of the model, and the *online recognition* to classify new observations. Classification is instantaneously based on the RLS score, responding with a vote either of biological class if the score is positive or of non-biological class when negative.

To partially correct instability of the final classification due to temporary failures, votes are collected into a temporal buffer of 15 frames and labels are attributed when at least the 60% of the votes is for one of the two classes, otherwise the system notifies the temporary uncertainty of the feedback. The **BioMerger** module synchronizes the feedbacks from the two classifiers and prepares a topographic feature

[1] Our solution accounts for a generic number N of moving entities in the scene, however, without loosing in generality, we focus on the case $N = 2$ to evaluate it in a controlled scenario.

map designed to compete with other feature maps in the visual attention system **PROVISION** (PROactive VISion attentiON) (Rea, Sandini, and Metta 2014), which generates a saccade command in correspondence to an attentive redeployment. Once the execution of the oculomotor action is completed the center of the camera (fovea of the robot eye) is relocated on the winning stimulus in the competition between perceptual features. For a more detailed description of the implemented modules see (Vignolo et al. 2016b).

Results

Data set

We acquired indoor videos of three subjects observed by the robot eyes while performing repetitions of given actions from a repertoire of dynamic movements typical of an interaction setting. More in details, we consider *Rolling dough* (9 movements, ∼300 frames), *Pointing* a finger towards a certain 3D location (7 movements, ∼330 frames), *Mixing in a bowl* (29 movements, ∼190 frames), *Transporting* an object from and to different positions on a table (6 movements, ∼300 frames), and *Writing* on a paper sheet (3 movements, ∼300 frames). As for the non-biological counterpart, we consider videos of a *Wheel with a random pattern* (∼300 frames) and of a *Wheel with a zig-zag pattern* (∼300 frames), a *Baloon* (300 frames), a *Toy Top* turning on a table (∼300 frames), and a *Toy Train* (∼398 frames). We acquired two videos for each non-biological and biological dynamic event. Henceforth, we will adopt the notation $\{V_{S_i}1\}$ and $\{V_{S_i}2\}$, $i = 1, 2, 3$, 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 depicting non-biological events. The images have size 320×240 and have been acquired at an approximate rate of 15 fps.

Offline Validation

To measure the capability of our model to generalize to new scenarios with respect to the training set, we trained the model using as training set a subset of the collection of videos from three subjects and of the non-biological dynamic events, i.e. $\{V_{S_1}1\} \cup \{V_{S_2}1\} \cup \{V_{S_3}1\} \cup \{V_{N1}\} \cup \{V_{N2}\}$. As a multi-resolution schema for the model, we adopted the combination that concatenates the raw features vector with the filtered measures on temporal windows $w_T = 15$ and $w_T = 30$, with a final feature vector of length 12 (see (Vignolo et al. 2016a) for the validation of the selection).

For testing, we analyzed system performance in scenarios of increasing complexity. In particular, we tested its robustness in classifying actions not belonging to the training set, actions performed by different subjects and hence characterized by different kinematics properties and non-biological motions characterized by different trajectories and velocities. Moreover, we also included particularly challenging scenarios, as the observation of *shadows* of trained and novel actions and actions performed in presence of *occlusions* (see caption of Figure 2 for a detailed description).

Figure 2 gives an impression of the overall classification results. Cases I and II, referring to biological scenarios, are very appropriately handled, with accuracies well above 90%. In Case III the method shows robustness with respect to new human actions, speaking in favor of its capabilities in capturing the regularities of human motion; in Case IV the presence of new subjects seems to influence the results. Cases V and VI 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. The two final cases consider non-biological dynamic events. As for Case VII, we may observe that the change in velocity profile of a known event is nicely accommodated by our model; in Case VIII (that we may consider in fact as instance of unknown non-biological dynamics, since both velocity profiles and spatial trajectories are subject to variations) the performances remain very satisfying.

Online Validation on the Robot

We first train the intelligent system with a set of biological categories (*Rolling dough, Pointing, Mixing, Transporting, Writing, Waving hand*) and a set of non-biological events (*wheel with random and zig-zag patterns, balloon, toy top and toy train*). On average, each video lasts about $20''$. Training is performed from completely blank a-priori knowledge meaning that, before training, the intelligent system lacks of the abilities of discriminating between biological and non-biological motion. The training is performed online replicating the situation where the operator interactively supervise the training. Model selection is also performed online.

We test the classification system by proposing a biological and a non-biological movement (distractor) in different portions of the iCub field. Our module, interfacing with PROVISION and Gaze Control System (Pattacini et al. 2010), should guide the proactive vision system to fixate the biological movement, bringing it to the center of the image plane. We measure the distance between the location of the biological stimulus on the image plane, provided by a color segmentation module as ground truth, and the center of the image, after the saccade. Table1 reports the results of the online validation. All the responses converge to a mean error in the range [20-40] pixels, corresponding to a metric range [4-8] cm, given the distance of the camera from the stimuli (64cm). This distance is reasonable, as an error of 40 pixels is correctly interpreted as a correct saccade from a human observer. More details on the results can be found in (Vignolo et al. 2016b).

Conclusions

In the proposed research, we investigate computational models of the visual primitives that are at the basis of social interaction in humans. Our inspiration roots on the very first stage of development, where the limited amount of visual information suggests that human beings have the capability to accomplish simple pro-social tasks on the basis of rather coarse motion models. We took inspiration from the Two-

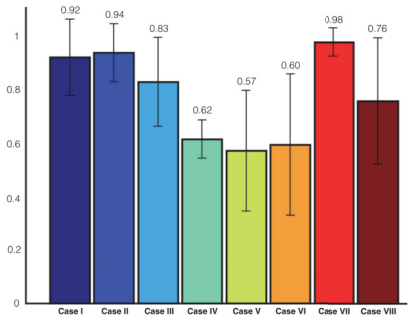

Figure 2: Overall classification accuracy of the model when evaluated on new test scenarios. In detail:
Case I: same conditions of the biological movements of the training set, using the second videos of each subject, i.e. $\{V_{S_1 2}\} \cup \{V_{S_2 2}\} \cup \{V_{S_3 2}\}$.
Case II: the three training subjects performing faster training actions (*Rolling dough* and *Transporting*).
Case III: one of the training subjects performing different actions classes (*Lifting* an object, *Gesticulating* while talking, and *Waving*) in different rooms and times.
Case IV: as Case III, considering different actions classes (*Lifting* an object, *Gesticulating* while talking), but observing a subject not considered in the training set.
Case V: a training subject performing actions included in the training set in and a new one (*Waving*) with occlusions.
Case VI: observing the shadow of an action included in the training set (*Pointing*) and a new one characterised by a whole-body motion (*Walking*) as opposed to the upper-body motions considered in the training set.
Case VII: the *wheel* with same patterns of the training set and a new one, with slower or faster rotation.
Case VIII: the *Toy train* covering a circular trajectory as opposed to the ellipsoidal path considered in the training set, with slower and faster velocity profiles (at approximately, respectively, half and twice the velocity of the training set). (see text for details on the different cases). From (Vignolo et al. 2016a)

.

Thirds Power Law, validating its applicability to video analysis problems. We demonstrated the possibility to exploit our method to perform human activity detection also in complex scenarios, where traditional shape-based approaches (e.g., skin or face detection) would fail. Our approach is robust to severe occlusions or to indirectly representation of the agent motion in the scene (as during the observation of agents' shadows). Moreover, we demonstrated the feasibility to engineer an online version of the method on a robotic intelligent system, which leverage the human detection skill and appropriately orient the focus of attention in order to establish an interaction with the human counterpart. These results represent the first step in the design of a hierarchical framework replicating the developmental stages of human visual perception and supporting social intelligence. By building on this capability of recognizing biological motion as proxy for the localization of interactive partners, we are now focusing on the capability of understanding classes of actions in order to prepare the interaction.

Table 1: Online Validation Results

Stimuli (L-R)	Perception acc.	corr/tot
gesturing-wheel random	27.30 ± 7.91	11/11 sac.
leaves-**writing subject1**	19.27 ± 7.83	10/11 sac.
cars-**gesturing**	22.98 ± 8.88	15/15 sac.
bouncing ball-**mixing**	20.95 ± 7.78	11/12 sac.
mixing,no person-wheel zigzag	25.02 ± 17.70	11/11 sac.
wheel random-**writing subject2**	25.53 ± 5.85	13/13 sac.

Acknowledgments

This research has been conducted in the framework of the European Project CODEFROR (FP7-PIRSES-2013-612555).

References

Farnebäck, G. 2003. Two-frame motion estimation based on polynomial expansion. In *Proc.s of the 13th Scandinavian Conference on Image Analysis*, SCIA'03, 363–370.

Lacquaniti, F., and Terzuolo, C. 1983. The law relating the kinematic and figural aspects of drawing movements. *Acta Psychologica* 54:115–130.

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(1):e186.

Metta, G.; Natale, L.; Nori, F.; Sandini, G.; Vernon, D.; Fadiga, L.; Von Hofsten, C.; Rosander, K.; Lopes, M.; Santos-Victor, J.; et al. 2010. The icub humanoid robot: An open-systems platform for research in cognitive development. *Neural Networks* 23(8):1125–1134.

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 *Intelligent Robots and Systems (IROS), 2010 IEEE/RSJ International Conference on*, 1668–1674.

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*.

Simion, F.; Regolin, L.; and Bulf, H. 2008. A predisposition for biological motion in the newborn baby. *Proceedings of the National Academy of Sciences* 105(2):809–813.

Tacchetti, A.; Mallapragada, P. K.; Santoro, M.; and Rosasco, L. 2013. Gurls: a least squares library for supervised learning. *The Journal of Machine Learning Research* 14(1):3201–3205.

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*.

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 of Humanoids Robotics*.