Preprint version; final version available at link.springer.com
Living Machines (2012), pp: 192-203
Published by: Springer
DOI: 10.1007/978-3-642-31525-1_17

Towards contextual action recognition and target localization with active allocation of attention

Dimitri Ognibene, Eris Chinellato, Miguel Sarabia, and Yiannis Demiris

Imperial College London, South Kensington Campus, London SW7 2AZ, UK
{d.ognibene,e.chinellato,miguel.sarabia,y.demiris}@imperial.ac.uk

Abstract. Exploratory gaze movements are fundamental for gathering the most relevant information regarding the partner during social interactions. We have designed and implemented a system for dynamic attention allocation which is able to actively control gaze movements during a visual action recognition task. During the observation of a partner's reaching movement, the robot is able to contextually estimate the goal position of the partner hand and the location in space of the candidate targets, while moving its gaze around with the purpose of optimizing the gathering of information relevant for the task. Experimental results on a simulated environment show that active gaze control provides a relevant advantage with respect to typical passive observation, both in term of estimation precision and of time required for action recognition.

Keywords: active vision; social interaction; humanoid robots; attentive systems; information gain

1 Introduction

The introduction of active vision [2, 1] was a fundamental step towards overcoming the limits of the classical vision paradigm as formulated by Marr [11]. Nevertheless, the perception of dynamic events still poses fundamental problems, such as the timely detection of the relevant elements, and the recognition of the discriminant dynamics. An archetypal and behaviorally relevant example of event perception is the recognition of an action executed by another agent. In order to deliver a behavioural advantage, and allow for timely action selection, the target and the end effector of an action should be predicted in advance, notwithstanding the limited perceptual and computational resources of the observer, and its knowledge of the environment, which is never optimal, due to occlusions and inner visual complexity.

We present here an attention system which integrates top-down with bottom-up attentional mechanisms. Starting from the simulation theory of mind point of view for action perception, we manage attention allocation in an active way, according to the predicted plausibility of candidate actions and possible targets. For a given action, the information that the attention system extracts during action observation is the state of the variables that the corresponding inverse

Preprint version; final version available at link.springer.com
Living Machines (2012), pp: 192-203
Published by: Springer
DOI: 10.1007/978-3-642-31525-1_17

model would control if it was executing that same action. For example, the inverse model for executing an arm movement will request the state of the arm when used in perception mode. This novel approach provides a principled way for supplying top-down signals to the attention system, which is to be integrated with bottom-up signals such as saliency maps or movement detectors. The influence of different attention biases can be modulated according to the task, the perceived interaction stage, what we know regarding the partner, and so forth.

We consider top-down attention as a competition of resources between multiple inverse models that seek to confirm their hypotheses about what the demonstrator's action/intention is. The saliency of a request for resources from each inverse mode can be linked to the quality of the predictions it offers. The computational and sensorimotor resources of the robot are distributed to the different inverse models as a function of the quality of the predictions they offer about forthcoming states of the interaction. In the meanwhile, a continuous estimation of environmental affordances allows for a dynamical update of which inverse models are applicable to the current state of the interaction.

In this way, the system is able to provide a prediction of the position of the observed agent effector, and thus an interpretation of his action, and also an estimation of the location of objects in the environment which constitute potential targets for the action being executed. Saccadic movements are performed according to a certain confidence level attributed to each of the competing models, and to the saliency of a feature (either hand or object).

Differently from previous approaches, which required the knowledge of the features of the different targets present in the environment to detect them [5], or the knowledge of their positions, in this work we propose a model that can overcome these limits, allowing for simultaneous exploration of the environment and recognition of the actions, exploiting both source of information to achieve faster action recognition. Also, according to a foveal model of vision, we consider that visual information gets more reliable and less noisy moving from the periphery to the center of the visual field.

2 The problem

Perception in active vision is constituted by a sequence of visual shots interleaved by saccadic movements[1, 2], aimed at purposefully exploring the environment, in order to extract the information relevant for pursuing the current goals. Given this strategy, the quality of the obtained information is due in great part to efficient and intelligent gaze control. A fundamental issue on this regard is the implicit indetermination of attending to something we cannot precisely locate yet. This requires a concurrent evolution of both the knowledge regarding the environment and the quality of the attention strategy. In our case, for achieving a shared, dynamical attention allocation during a social interaction, decisions on where to look are strictly linked to the movements of the partner. This adds further complexity to the task, which now has to account for a changing visual scenario. Human behavioural studies, on tasks like face recognition [7] and visuo-

Preprint version; final version available at link.springer.com
Living Machines (2012), pp: 192-203
Published by: Springer
DOI: 10.1007/978-3-642-31525-1_17

motor control [15, 10], have shown that humans are able to adapt their visual exploration to the specific requirements of the task at hand.

Robotic studies on adaptive active vision so far have focused on the previously mentioned topics in isolation. In [17] an artificial fovea is controlled by an adaptive neural controller. Without a teacher, this learns trajectories causing the fovea to find targets in simple visual scenes and to track moving targets. The model in [9] solve active sensing problems under uncertainty. A reinforcement learning algorithm allows it to develop active sensing strategies to decide which uncertainties to reduce. However, in this study the model of the task is known a-priori and motor control is hardwired. Other works (e.g. [3, 19]) employ evolutionary learning techniques for developing adaptive active vision systems. These approaches are robust to the *perceptual aliasing problem*, however they do not allow on-line adaptation to changing environments. In [12] a neural architecture for eye arm coordination is proposed which learns autonomously task-specific attentional policies, exploiting a strong link between attention and execution of actions. The authors also proposed that a bottom-up attention system can be exploited to bootstrap learning, and hypothesised on the basis of neural simulations that the limited size of fovea can play an important role in the efficiency of learning [14].

In this work, we deal with the above issues by letting an integrated attention system assume gaze control while observing a partner performing a reaching action toward one of a small set of target objects. Neither the goal of the action, nor the exact location of the potential targets are known beforehand, so that hand trajectory and target position have to be estimated contextually while trying to understand what is the action goal, i.e. where the partner is moving its hand towards. An example of a possible experimental setup is provided in Fig. 1, where the humanoid robot iCub is observing a human partner starting a reaching movement towards one of three potential target objects placed in the common working space. The robot has to decide where to observe (estimated hand or object position) in order to 1) estimate the objects exact positions and 2) understand where the partner's hand is reaching at.

3 Action recognition with dynamic allocation of attention

In this work, we build on some of the concepts introduced with the HAMMER model for action perception and imitation based on the direct matching hypothesis [4]. The system described here is based on the integration of the latest HAMMER framework implementation [16] with a gaze controller which directs attention in order to maximise discrimination performance, while maintaining robustness to noise, and a contextual estimation of both end effector location and position of all potential targets.

A number of different models, at least one for each of the possible targets of the action, concur for both attention allocation and for the final discrimination of the action goal. Following HAMMER guidelines [5], the discrimination between the available action hypotheses is based on the computation of a confidence

Preprint version; final version available at link.springer.com
Living Machines (2012), pp: 192-203
Published by: Springer
DOI: 10.1007/978-3-642-31525-1_17

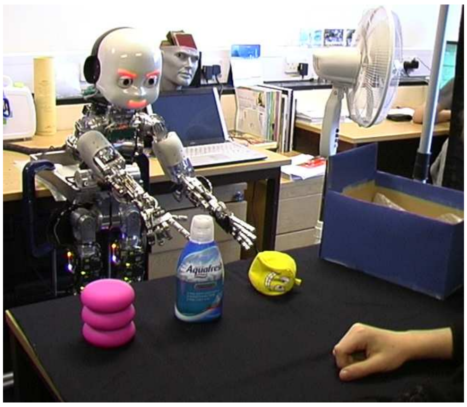


Fig. 1. Example of experimental setup, with iCub looking at target objects and arm movement (its own pointing movement is not relevant here).

value that measures the overall Euclidean distance between the predicted action trajectories and the observed motion trajectories. To compute such prediction, HAMMER uses a combination of forward and inverse model pairs which are the same models that can be used for action control.

Formally, an inverse model is a function that, given a certain goal g, maps the current state S to the action A the agent has to execute to achieve the goal: $i_g : S \to A$. A forward model is a function that maps the current state and the action being executed to the next expected state $f_g : S\ timesA \to S$.

In this work, the candidate actions among which the observer has to chose are different reaching movements toward different targets in space, g. A reaching model m_g, composed of a pair of inverse and forward models (i_g, f_g) is required for each different g. Each reaching model works directly in the space of the end-effector, and the action space is coincident with the state space ($A \equiv S$), because the used inverse model computes the next desired end-effector position $\mathbf{p}^{t+1}$, and the forward model returns the same value, too.

The following is the equation, akin to a PID controller, employed by a model m_g to compute the next position $\mathbf{p}^{t+1}$, when the target is at position $\mathbf{p}_g$:

$$\mathbf{p}^{t+1} = \mathbf{p}^t + \tau\{\dot{\mathbf{p}}^t + \tau[K(\mathbf{p}_g - \mathbf{p}^t) - D\dot{\mathbf{p}}^t]\}. \tag{1}$$

This equation leads to a motion with a smooth linear trajectory that brings asymptotically toward the target. The confidence function for each model and time step c_t^g is updated employing the difference between the predicted end-effector position $\tilde{\mathbf{p}}^{t+1}$ and the perceived one $\mathbf{p}^{t+1}$:

$$c_{t+1}^g = \frac{1}{1.0 + \|\tilde{\mathbf{p}}^{t+1} - \mathbf{p}^{t+1}\|} + c_t^g. \tag{2}$$

For increased plausibility, we assume that the observations of the end-effector and of the affordances are affected by noise that is dependent on the sensors

Preprint version; final version available at link.springer.com
Living Machines (2012), pp: 192-203
Published by: Springer
DOI: 10.1007/978-3-642-31525-1_17

configuration, i.e. gaze position. Thus, if the observer gaze position is $\mathbf{pos}_o$, the actual observation of an object at position $\mathbf{p}$ is distributed according to:

$$\mathbf{N}(\mathbf{p}, 0.15\|\mathbf{p} - \mathbf{pos}_o\|^2). \tag{3}$$

This noise model is an approximation of human foveal vision, where the most of the visual receptors are located in the central area (fovea) of the retina and their density, and thus the visual resolution, decreases departing from the fovea.

In order to manage the noisy input, each model uses Kalman filters for the estimation of the end-effector and affordance positions, and an active vision system is integrated that exploits the estimation of the uncertainty produced by the Kalman filters. The main assumption is that the observer can use features which allow to discriminate between the different affordances and the effector. This approach is similar to that of [8] and [18] but, instead of being limited to track or to find objects in a dynamic environment, it allows for the active recognition of a dynamic event.

In this work, independent Kalman filters are used for each action element. An action element has position $\mathbf{p}$ and produces an observation z. The associated Kalman filter produces an estimated probability distribution $b_\mathbf{p}$ and a corrected probability $\hat{b}_\mathbf{p}$, which uses the observation received at the current time step.

In order to produce these estimations the Kalman filter uses a process model and noise model. The process model has the form $\mathbf{p(t+1)} = \mathbf{Ap}(t) + \mathbf{b}(t)$. For both the effector and the targets, matrix $\mathbf{A}$ is the identity matrix $\mathbf{I}$. We assume that the targets are still, thus their process noise $\mathbf{b}(t)$ is zero. The effector process noise is 0.01 $\mathbf{I}$ to model small changes in the trajectories.

In this work, we decoupled the prediction and the correction phases of the Kalman filters for the end-effector in each model. The mean position of the end-effector is updated in accordance with the action model in eq. 1, after the correction of estimated position of the target and the prediction (not corrected) of the end-effector using the process model as described above. The variance estimated by the process model is not modified. After this update the Kalman correction phase takes place also for the end-effector.

In our task, we need to take into account the implicit imprecision of the sensory information, together with the lack of exact knowledge regarding hand and targets position. As a consequence, the typical Kalman formulation have to be adapted, so that the observation models of the Kalman filters are able to account for the change of the sensory configuration and the uncertainty regarding the real environment. While observation noise increases with the distance between observation point and real object position, the latter is not known, and only a prior estimation $b_\mathbf{p} = \mathbf{N}(\bar{\mathbf{p}}, \mathbf{\Sigma}_p)$ is available before saccade execution. Thus, the resulting observation model depends on current gaze position and belief state:

$$P(\mathbf{z}|\mathbf{pos}_o, b_\mathbf{p}) = \int p(\mathbf{z}|\mathbf{p}, \mathbf{pos}_o) b_\mathbf{p} d\mathbf{p}. \tag{4}$$

Preprint version; final version available at link.springer.com
Living Machines (2012), pp: 192-203
Published by: Springer
DOI: 10.1007/978-3-642-31525-1_17

The implemented observation model is expressed by the following normal distribution, considering that Kalman filters assume Gaussian distributions and linear dynamics:

$$P(\mathbf{z}|\mathbf{pos}_o, b_\mathbf{p}) \approx \mathbf{N}(\bar{\mathbf{p}}, 0.15(\|\bar{\mathbf{p}} - \mathbf{pos}_o\|^2\mathbf{I} + 0.9\mathbf{\Sigma}_p)). \tag{5}$$

The attention system uses the probability distribution estimated by the Kalman filters in all the models and the confidence value of each action hypothesis. The attention system minimises uncertainty on the most probable action. The effects of reducing uncertainty between the different action hypotheses is not directly taken into account by the current system for computational reasons. Gaze point selection currently considers only instantaneous saccades even if the system is allowed to select a new saccade target only after the previous attentive action has been completed.

Each element of each model, e.g. target and effector , is considered independently by the attention system instead of integrating the different probability distributions associated to elements shared by different models. e.g. the different expected positions of the end-effector for the different models. In this implementation each action hypothesis has two elements, effector and affordance, and the attention system selects targets from a set of $2 * n_a$ elements, where n_a is the set of action hypotheses.

The selected target, with estimated position $\bar{\mathbf{p}}$ related to hypothesis g, is the one which maximizes the following objective function:

$$\log(|\mathbf{\Sigma}_p|)(1 + c^g). \tag{6}$$

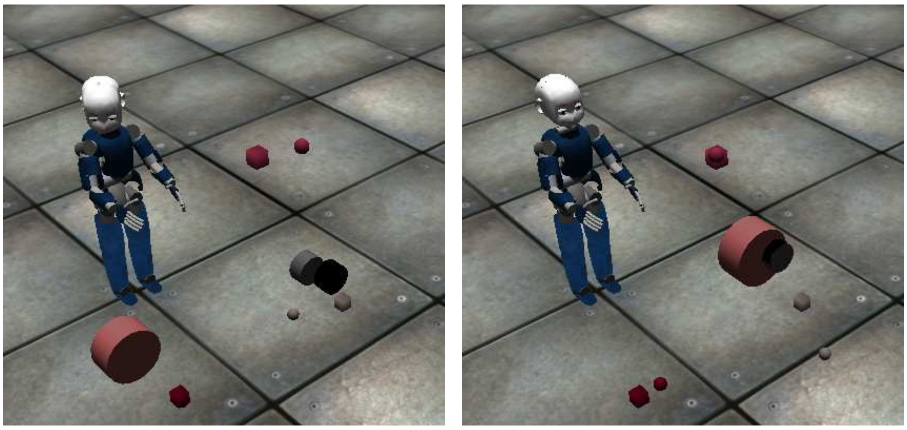
This objective function accounts at the same time for both the reduction of uncertainty, which can be measured using entropy (i.e., in the case of a Gaussian distribution, by the logarithm of the determinant of the covariance matrix), and the relevance for the most probable action hypothesis.

There are several approximations in this objective function: a) it does not consider the residual entropy of the target assuming that after the saccade the object will be perfectly centered; b) it does not consider the information gain on the other targets. At the same time, this formulation allows to select only positions corresponding to estimated targets, while other positions may allow to increase the overall information gain.

4 Experimental evaluation

The proposed model has been implemented on the iCub Simulator where the simulated robot head was controlled by the attentional system. In the simulated environment (Fig. 2) three target objects were created (small coloured boxes). The system receives noisy observations, represented as small spheres of the same colour as the actual objects. The simulated sensor noise is proportional to the square of the distance between the real position (boxes) and the robot gaze point (red cylinder), following Eq. 3. The end-effector of the other agent is displayed

Preprint version; final version available at link.springer.com
Living Machines (2012), pp: 192-203
Published by: Springer
DOI: 10.1007/978-3-642-31525-1_17



(a) Fixating laterally (b) Fixating at the hand

Fig. 2. Experimental setup with example of different gaze points. The big red cylinder represents the gaze point of the robot. The boxes represent the real position of the observed action target while the small spheres represent the related observations that the robot senses. The dark small cylinder represents the end effector that is executing the action and the brighter one the related observation.

by a black cylinder, while its observation is a gray cylinder. The head and eye are controlled using the fixation point and the *iKinGazeCtrl* iCub module[1]. A new fixation point is sent only when the previous movement has ended.

The end-effector of the other agent is moving, according to equation 1 with $K = 3$ $D = 0.5$ and $\tau = 0.04$, towards one of the target object, randomly selected. The system is provided with three models, with the same parameters for the three different objects, and with prior $\mathbf{N}(o, I)$, where o is an observation sampled according to the noise model, constituting the initial gaze point. Examples of hand trajectories devised by the concurrent models are depicted in Fig. 3. It can be observed how the hand terminates in each of the three target objects corresponding to the three candidate models. The top trajectory is the one which is actually performed in this case. In order to characterise the behaviour of the system in different working conditions and with different setup and parameters, a number of analyses can be performed on the system performance. The results presented in this section take into account different aspects of the system behaviour, and are currently being employed to improve the reliability and generalization skills of our action recognition module.

First of all, we need to verify whether exploratory gazing movements performed according to the attention control provide an actual advantage with respect to a typical passive perception paradigm. We have performed experiments with moving gaze, performed as described above, and with steady gaze, in which the robot was fixating the same random location of its working environment for

[1] http://eris.liralab.it/iCub/main/dox/html/group_iKinGazeCtrl.html

Preprint version; final version available at link.springer.com
Living Machines (2012), pp: 192-203
Published by: Springer
DOI: 10.1007/978-3-642-31525-1_17

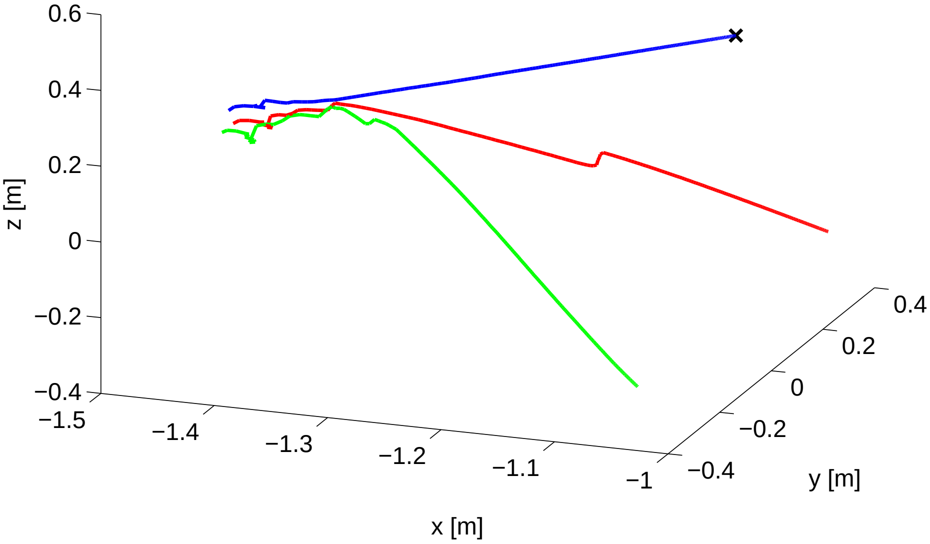


Fig. 3. Estimated hand trajectories according to the three competing models, which converge asymptotically to the three targets. Sudden changes in the two lower estimated trajectories are due to gazing actions that change the related observation models.

the whole test. In both moving and steady gaze experiments we made sure that all relevant stimuli were always visible to the robot. We executed twenty trials with each of the paradigms, and averaged their outcomes, obtaining the results summarized in Table 1. **Max confidence error** is the overall mean distance between the real effector position and the estimation provided by the most credited model (i.e. the one with highest confidence) at each time step. **Winner error** is the same mean distance computed at each time step for the winning model (this error can only be computed *a posteriori*, when it is known what model has prevailed). There is a clear difference between the methods in this regard, as the moving gaze paradigm error is about half of the steady gaze error. It is also interesting to observe that, in both paradigms, the two different errors assume similar values, indicating that the **Max confidence error** represents a good on-line approximation of the actual performance of the winner model.

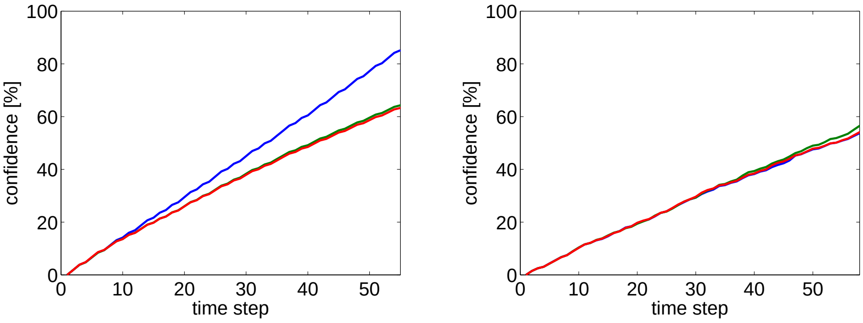
Indexes **Choice TS** in Table 1 represent potential decision moments, according to two different thresholds. More exactly, they show the time step at which the confidence of the dominant model, computed according to Eq. 2, is 20% and 10% higher than the others, respectively. The performance of the attention-based protocol is again clearly better than the passive protocol. The 20% threshold is achieved more than 5 time steps earlier on average by the former (corresponding to a 10% improvement). Even more significantly, the 10% threshold is attained about 8 time steps earlier on average, which is like saying that the moving gaze system decides 24% more quickly than the steady gaze system.

Preprint version; final version available at link.springer.com
Living Machines (2012), pp: 192-203
Published by: Springer
DOI: 10.1007/978-3-642-31525-1_17

Table 1. Comparison of performance in trials with moving and fixed gaze

	Moving gaze	**Steady gaze**
Max conf. error	0.049	0.118
Winner error	0.048	0.099
Choice TS (20%)	44.6	49.9
Choice TS (10%)	24.4	32.1

A typical evolution of the confidence level for each of the three competing models can be observed in Fig. 4 for both Moving gaze and steady gaze protocols. It is again possible to observe how the active paradigm is able to differentiate the goal action around time step 20 (Fig. 4(a)), whilst the passive paradigm seems to fail completely in the task (only at the very end of the trial a small, still non-significant prevalence of one of the models can be spotted, Fig. 4(b)).



(a) Moving gaze

(b) Steady gaze

Fig. 4. Evolution in time of confidence with (a) and without (b) attentional gaze control. With gaze control the right action is clearly identified before time-step 20. Without gaze control the agent is not able to recognize the performed action.

To better understand how the active exploration of the environment through the execution of saccadic movements is performed, Fig. 5 shows an example of the evolution of gaze direction during an experiment, computed by the attention model as described in the previous section. Three different phases are highlighted. During the first half of the trial (dots in Fig. 5), gaze moves rather erratically all around the task space, but after this bootstrapping phase more regular behaviours can be observed. Time steps from half to three quarters of action execution show gaze points approximately distributed along the dominant hand trajectory, suggesting that the system has now understood where the action is going on (plus symbols in Fig. 5). Finally, in the last quarter of the trial, one of the model seems to be clearly dominant over the others. Estimations of both hand trajectory and location of target object are reasonably accurate, and

Preprint version; final version available at link.springer.com
Living Machines (2012), pp: 192-203
Published by: Springer
DOI: 10.1007/978-3-642-31525-1_17

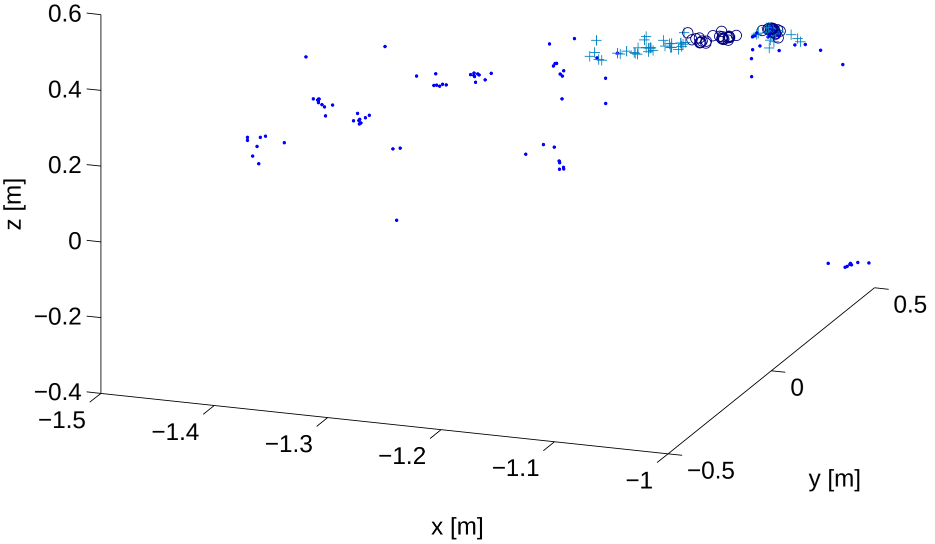


Fig. 5. Evolution of gaze point during different stages of action observation (dots: first half; plus: third quarter; circles: fourth quarter).

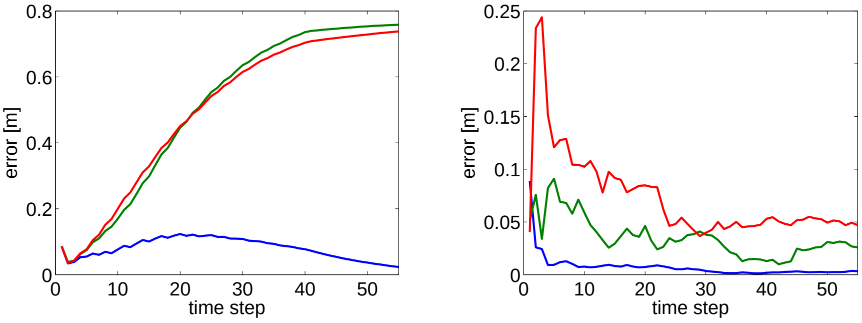
the dominant model makes the system move forth and back between these two locations, which are now definitely considered the most interesting, to further improve their estimation (circles in Fig. 5).

The last results we present concerns the actual prediction capabilities of the system in terms of estimation of hand trajectory and object position. Fig. 6 shows the error observed in the approximation obtained by each model in its estimation of the effector trajectory and target location. It can be observed that only one of the models achieves a correct estimation of the actual hand position (i.e. the model that correctly recognize the action), whilst the others wrongly convergence towards the wrong targets (Fig. 6(a)). Nevertheless, Fig. 6(b) shows that even the targets associated to the losing models are detected with a good approximation, and the first of the phases depicted in Fig.. 5 is critical in this regard, as it allows to achieve a good representation of all stimuli in the environment while gradually shifting the focus toward the supposedly most interesting one.

5 Conclusions

The results reported in this work show that the proposed approach is viable for the problem of action recognition in unknown environments. A relevant contribution given by this paper is related to the importance of using the simulation approach when perceiving actions with limited perception. Differently from the teleological and associative approaches (see [13] for the distinction), the simulation approach describes mechanisms that implement action recognition by producing dynamic internal representations that can be used also to direct at-

Preprint version; final version available at link.springer.com
Living Machines (2012), pp: 192-203
Published by: Springer
DOI: 10.1007/978-3-642-31525-1_17



(a) Hand estimation error

(b) Target estimation error

Fig. 6. Evolution of estimation error on hand position (a) and on target position (b) according to the three competing models. End-effector position is estimated correctly only by one of the models (a), while target location is estimated with good approximation by each model, reaching a 50mm error in the worst case.

tention. This is particularly evident in this model, which actually employs the simulated position to drive attention, whilst in previous models prediction was used to modulate bottom-up attention (see e.g. [6]). Experimental results have confirmed the advantages of the attention-based action recognition system, both in terms of precision and, maybe more importantly, in terms of decision time. This latter aspect is indeed critical when an agent has to interpret or recognise a partner's action. For a robot, being able to understand what a human partner is doing some tenth of a second earlier can be of fundamental importance in order to achieve a meaningful interaction, avoiding the inconvenient obligation for the human to wait for the robot to interpret his movements.

We are now working on the recognition of real human actions, and the comparison with human performance in the same task. We aim to substitute the current action models with more realistic ones that account for more peculiarities of human movements, that can thus allow for higher accuracy and faster recognition, i.e. using the pre-shaping of hand for gasping. Then, we plan to test the robustness of the system with respect to incorrect models, and to the presence of a high number of action hypotheses, differentiated also for the different parameters chosen, e.g. different execution speeds. Another interesting improvement would be the use non myopic target selection, for a better estimation of the relative importance of targets and effectors in visual perception of actions.

Acknowledgments

This research has received funding from the European Union Seventh Framework Programme FP7/2007-2013 Challenge 2 Cognitive Systems, Interaction, Robotics under grant agreement No [270490]- [EFAA].

Preprint version; final version available at link.springer.com
Living Machines (2012), pp: 192-203
Published by: Springer
DOI: 10.1007/978-3-642-31525-1_17

References

1. R. Bajcsy. Active perception. *Proceedings of the IEEE*, 76(8):966–1005, 1988.
2. D.H. Ballard. Animate vision. *AI*, 48:57–86, 1991.
3. G.C.H.E. de Croon, E.O. Postma, and H.J. van den Herik. Adaptive gaze control for object detection. *Cognitive Computation*, 3(1):264–278, 2011.
4. Y. Demiris and B. Khadhouri. Hierarchical attentive multiple models for execution and recognition of actions. *Robotics and Autonomous Systems*, 54:361–369, 2006.
5. Y. Demiris and B. Khadhouri. Content-based control of goal-directed attention during human action perception. *Journal of Interaction Studies*, 9(2):353–376, 2008.
6. Y. Demiris and G. Simmons. Perceiving the unusual: temporal properties of hierarchical motor representations for action perception. *Neural Networks*, 19(3):272–284, Apr 2006.
7. J. J. Heisz and D. I. Shore. More efficient scanning for familiar faces. *J Vis*, 8(1):1–10, 2008.
8. K. Kastella. Discrimination gain to optimize detection and classification. *IEEE Transactions on Systems, Man and Cybernetics, Part A: Systems and Humans*, 27(1):112–116, 1997.
9. C. Kwok and D. Fox. Reinforcement learning for sensing strategies. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS2004)*, 2004.
10. M. F Land. Eye movements and the control of actions in everyday life. *Prog Retin Eye Res*, 25(3):296–324, 2006.
11. D. Marr. *Vision: A Computational Investigation into the Human Representation and Processing of Visual Information.* W. H. Freeman, New York, 1982.
12. D. Ognibene, C. Balkenius, and G. Baldassarre. Integrating epistemic action (active vision) and pragmatic action (reaching): A neural architecture for camera-arm robots. In *Proceedings of the Tenth International Conference on the Simulation of Adaptive Behavior*, 2008.
13. D. Ognibene, Y. Wu, K. Lee, and Y. Demiris. Hierarchies for embodied action perception. Under review, 2012.
14. D. Ognibene, G. Pezzulo, and G. Baldassarre. How can bottom-up information shape learning of top-down attention control skills? In *Proceedings of 9th International Conference on Development and Learning*, 2010.
15. U. Sailer, J. R. Flanagan, and R. S. Johansson. Eye-hand coordination during learning of a novel visuomotor task. *J Neurosci*, 25(39):8833–8842, 2005.
16. M. Sarabia, R. Ros, and Y. Demiris. Towards an open-source social middleware for humanoid robots. In *Proc. 11th IEEE-RAS Int Humanoid Robots (Humanoids) Conf*, pages 670–675, 2011.
17. J. Schmidhuber and R. Huber. Learning to generate artificial fovea trajectories for target detection. *Int J Neural Syst*, 2(1-2):135–141, 1991.
18. E. Sommerlade and I. Reid. Information theoretic active scene exploration. In *Proc. IEEE Computer Vision and Pattern Recognition (CVPR)*, May 2008.
19. M. Suzuki and D. Floreano. Enactive robot vision. *Adapt Behav*, 16(2-3):122–128, 2008.