Teaching a Humanoid Robot to Recognize and Reproduce Social Cues

Conference Paper · October 2006

DOI: 10.1109/ROMAN.2006.314458 · Source: IEEE Xplore

Sylvain Calinon
Idiap Research Institute
102 PUBLICATIONS 3,917 CITATIONS
SEE PROFILE

Aude Billard
École Polytechnique Fédérale de Lausanne
301 PUBLICATIONS 8,427 CITATIONS
SEE PROFILE

Teaching a Humanoid Robot to Recognize and Reproduce Social Cues

Sylvain Calinon and Aude Billard*

***Abstract*— In a *Robot Programming by Demonstration* framework, several demonstrations of a task are required to generalize and reproduce the task under different circumstances. To teach a task to the robot, explicit pointers are required to signal the start/end of a demonstration and to switch between the learning/reproduction phases. Coordination of the learning system can be achieved by adding social cues to the interaction process. Here, we propose to use an imitation game to teach a humanoid robot to recognize communicative gestures, which then serve as social signals in a *pointing-at-objects* scenario. The system is based on Hidden Markov Models (HMMs) and use motion sensors to track the user's gestures.**

I. Introduction

Robot Programming by Demonstration explores novel means of teaching a robot new skills by showing directly to the robot how to perform a task. Such mechanism provides user-friendly means for the end-user to re-program the robot in a natural way, without any needs of programming/engineering skills. In previous work, we developed a probabilistic system capable of extracting the important characteristics of a task from multiple trials [1]–[3]. The system requires to produce different demonstrations of the same task, while collecting human motion data using motion sensors and vision. Then, the robot generalizes over the different demonstrations to reproduce the task under different situations. Thus, the robot can learn new skills in a fast and efficient way and reproduce them under new circumstances without the intervention of the user. However, explicit pointers are required to guide the scenario, i.e. to signal the start/end of a demonstration and to switch between the learning/reproduction phase. These explicit pointers are currently performed by pressing keys on a keyboard. This work aims at exploring possible solutions that could be used for coordinating our learning system, i.e. to conduct a simple interactive scenario between the user and the robot.

Humanoid robots are endowed with multiple sensors recording highly-dimensional multimodal signals and activated by complex redundant manipulators. They are faced with an incoming stream of data from which they must figure out what are relevant to a specific task. Imitation learning has been applied as an efficient way to narrow this search space and to estimate a control policy by observing the user's performance at resolving the task [4]. By extracting the redundancies and important characteristics of the task, the learning problem becomes significantly more tractable. In our work, the salient aspects of the task are determined by a probabilistic approach combining different *Machine Learning* tools [5]. These tools usually rely on many demonstrations to infer the task structure. In contrast, the availability of data is limited by the patience of the user. Here, we are looking for solutions to enhance the speed convergence of our statistical algorithms by exploiting additional social cues extracted during the teaching process. Indeed, explicit cues such as pointing gesture or gaze direction can guide the teaching process, by narrowing the search space for the selection of the important features to reproduce the task. Finally, the social interaction also aims at entertaining the user during the process (teaching the task), the interaction, and the result (executing the task once taught), see e.g. [6].

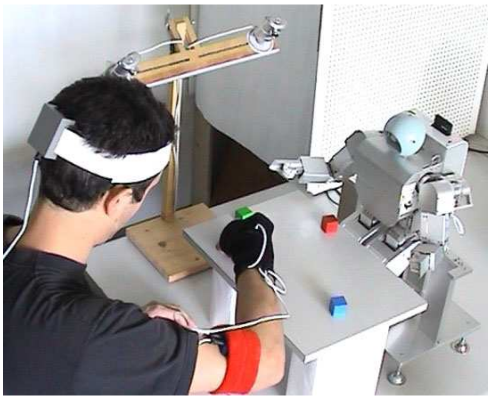

Fig. 1. Experimental setup. The user is pointing at an object while the robot is *observing* his gesture using motion sensors.

Several robots have been developed to explore the use of natural pointing and gazing cues to convey the intention of the user [7]–[13]. In the majority of these works, the user's gaze and pointing directions are extracted from cameras, with classifiers designed carefully to detect the occurrence of these cues. The natural pointing behavior is characterized by the gaze moving to the target first, and the arm/hand pointing to the target in a second phase, while maintaining the gaze to the target until the final arm/hand posture has been reached. During a goal-directed pointing movement, gazing preparation toward a new visual target is inhibited. Indeed, experiments showed that subjects were not able to initiate a saccade to a new target when the hand was reaching for a first target, i.e. subjects postponed the initiation of a new saccade until pointing was completed [14]. A well coordinated motion pattern can thus be observed during goal-directed arm movements, which can be encoded efficiently in a Hidden Markov Model (HMM). In this paper, we are interested in: 1) Testing a robust sensory solution based on

*S. Calinon and A. Billard are with the LASA Laboratory, EPFL, CH-1015 Lausanne, Switzerland {sylvain.calinon, aude.billard}@epfl.ch motion sensors to track the gestures of the user with a low-computational process, 2) Using HMM to learn automatically the essential user-dependent features of different communicative cues through an imitation game, 3) Incorporating the information extracted from these social cues to other statistical learning methods.

II. Experimental setup

A. Experimental scenario

The experimental setup and teaching scenario of the experiment are presented in Fig. 1 and 2, where the user brings the robot's attention to different objects on a table. During a first phase of the interaction, the user imitates the robot's behaviors, building a representation of the correspondence between its gestures and the user's gestures. The robot produces a gesture and observes the corresponding user's gesture. During a fixed time interval Δt, joint angles trajectories are collected from the motion sensors. The imitation game stops when the different gestures have been collected. During the second phase, the human uses this common understanding of basic behaviors to bring the robot's attention to locations of relevant objects. At each time step t, the signals collected during the time interval $\{t - \Delta t, t\}$ are compared to the different gesture models. When a pointing gesture is detected, information concerning gazing and pointing directions are collected until detection of a turn-yielding cue. At this time, the robot points at the object with highest probability and request an evaluation from the user. By shaking or nodding the head, the user can then point again at the object to clarify his/her selection or finish the scenario.

A preliminary calibration phase is performed offline by a user pointing at the corners of the table and at the different objects placed on this table, in order to initialize their position. The robot is then taught through *kinesthetic learning* how to point at the objects, and how to perform different communicative gestures. The robot is then provided with a database of gestures, but at the beginning of the scenario, it can not recognize any gesture and starts with only few *a-priori* on the communicative gestures (only a maximum time interval is defined). The robot learns the relevant features characterizing the different gestures through an imitation game played with the intervening user. Each person can then provide his/her own version of the gesture, that can have different characteristics than the one performed by the robot.

B. Motion sensors

In [15], we explored different means of conducting an interaction between a human user and a humanoid robot using speech and vision. Although these modalities are important *a-priori*, they require heavy computation and may be technically cumbersome. Expectations about these systems are often overestimated when compared to the human ability to process visual and vocal information, which can lead to dissatisfactory results considering the computer resource involved. In [5], we explored the use of motion sensors attached to the body of the demonstrator to convey information about human body gesture. Although these sensors are not directly related to human-like sensory abilities, they measure robust information about body posture, and can be used easily in different environment, independently of the sound, lighting and occlusion conditions. The main drawback is that these sensors must be attached to the body of the demonstrator[1].

In this work, we explore the use of *x-sens* motion sensors to process communicative gestures and to record human motion data for learning purpose. Gestures are recorded by 5 *x-sens* motion sensors attached to the torso, right upper-arm, right lower-arm, right hand and on the back of the head. Each sensor provides the 3D absolute orientation of each segment, by integrating the 3D rate-of-turn, acceleration and earth-magnetic field, at a rate of 100Hz with a precision of 1.5 degrees. A rotation matrix is defined as the orientation of a distal limb segment expressed in the frame of reference of its proximal limb segment. The kinematics motion of the different joints can then be computed by decomposing the rotation matrix into joint angles. Thus, 8 joint angles are recorded, corresponding to the degrees of freedom (DOFs) of our robot (1 DOF for the torso, 2 DOFs for the head, 3 DOFs for the shoulder and 2 DOFs for the elbow).

C. Gesture recognition

Communicative gestures are characterized by cultural and personal differences. Turn-taking and attention mechanisms can present subtle differences from one individual to another. Previous attempts at parameterizing manually the characteristics of the different communicative cues showed that the solutions require fine tuning and are often user-dependent. Here, we suggest to learn automatically these parameters through an imitation game, using Hidden Markov Models (HMMs). The robot can then robustly extract the characteristics of different head/arm gestures, and use these characteristics for recognition purpose, see e.g. [17], [18].

Although pointing and gazing cues are often related to static poses, useful information is contained in the establishment of these poses. It is supported by the literature on human development indicating that infants imitate facial expressions when the adult adopts the expression, but do not imitate when this expression is presented statically. Thus, the movement preceding a stable expression is a clue used by infants to notice a facial expression that is worth imitating [19]. Encoding of the temporal information is performed robustly by HMMs and allows to recognize a gesture even in presence of non-linear temporal distortion. The correlations between the different signals, along the motion, are also learned automatically by HMMs. It is relevant for pointing gesture because of the coupling between the gaze and the arm motion, see e.g. [20].

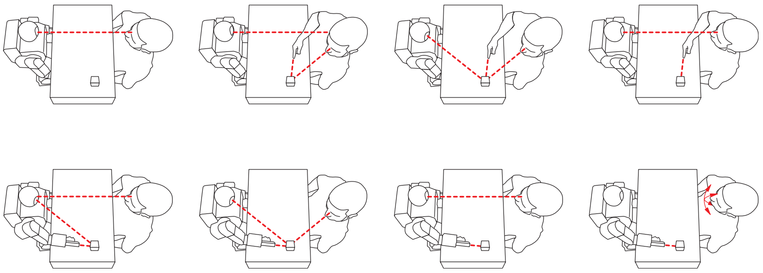

Fig. 2. Experimental scenario. *First line:* The user gazes at the robot to attract its attention (*turn-taking cue*). Then, he looks and points at an object in the environment. The robot follows his/her gaze, and *observes* the pointed object. The user gazes at the robot again (*turn-yielding cue*). *Second line:* The robot takes its turn, gazes and points at an object, while the user looks at the selected object. The robot gazes at the user again to request an evaluation of its selection (*turn-requesting cue*). Finally the user signals to the robot whether the correct object has been selected by nodding/shaking his/her head.

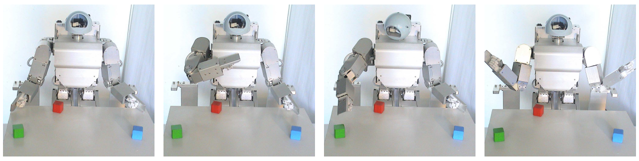

Fig. 3. Communicative gestures used by the robot (mutual gaze, turn-taking signal, object selection and request for clarification).

D. Humanoid robot

The experiments are conducted with a Fujitsu HOAP-2 humanoid robot with 25 DOFs, of which only 11 DOFs are used (2×4 DOFs for the arms, 1 DOF for the torso, 2 DOFs for the head). The remaining DOFs of the legs are set to a constant position, so as to support the robot in an upright posture, facing a table. In the experiments reported here, the robot is previously taught gestures through kinesthetics, i.e. by the demonstrator moving its arms and head. To achieve this, the robot's motors are set in a passive mode, whereby each limb can be moved by the human demonstrator. The kinematics of each joint are recorded at a rate of 1000Hz. The robot is thus able to perform several communicative behaviors, see Fig. 3.

III. Data processing

A. Hidden Markov Models

To avoid making assumptions on the spatio-temporal variability of the dataset, a fully-connected continuous HMM with full covariance matrix describing the output variables distribution is used[2]. Using such a model requires the estimation of a large set of parameters, which is optimally achieved when the dataset is large. In a *Programming by Demonstration* framework, the user should not have to produce more than a few demonstrations. This means that the set of parameters to learn is often quite large compared to the amount of training data. In the experiments presented here, we let the user choose the number of demonstrations that he wants to provide. If this number is below 5 (which is often the case), additional examples are generated and added to the training set, by adding Gaussian noise to the provided samples. Thus, even with a single trial, the system is still able to build a rough generalization of the gesture (i.e. by allowing a fixed variation on the data).

Expectation-Maximization (EM) algorithm is used to estimate the HMM parameters. It starts from initial estimates, and converges to the nearest local maximum of the likelihood function. Thus, initialization highly affects the model performance. To better estimate the state distribution of the HMM, we perform first a rough clustering of the data using *k-means*, as in [3]. Next, we estimate a *Gaussian Mixture Model* (GMM) by EM, using the *k-means* clusters at initialization. Finally, the dynamics, i.e. transitions across the states, are encoded in a HMM created with the GMM state distribution.

A dataset of N data of dimensionality D, $X = \{x_j\}_{j=1}^N$ with $x_j \in \mathbb{R}^D$, is modeled by a Gaussian mixture of K-components:

$$p(x_j) \;=\; \sum_{k=1}^{K} \pi_k \, \mathcal{N}(x_j; \mu_k, \Sigma_k)$$

where $\pi_k \in \mathbb{R}$ is the prior probability and $\mathcal{N}(x_j; \mu_k, \Sigma_k)$ is the D-dimensional Gaussian density of component k, with $\mu_k \in \mathbb{R}^D$ and $\Sigma_k \in \mathbb{R}^{D\times D}$ the mean and covariance matrix.

To determine the number of states in a HMM, heuristic methods are often used, sometimes not adequately tuned for HMM. In our approach, model selection is performed in the GMM initialization phase. Multiple GMMs are estimated, the best model is selected, and a single HMM estimation is performed. *Bayesian Information Criterion* (BIC) [22] is used to select the optimal number of components K:

$$S_{BIC} = -\mathcal{L} + \frac{n}{2}\,\log(N)$$

where $\mathcal{L}$ is the log-likelihood of the model, n is the number of parameters required for a mixture of K components, i.e. $n = (K-1) + K\left(D + \frac{1}{2}D(D+1)\right)$. N is the number of D-dimensional datapoints. The first term of the equation measures how well the model fits the data, while the second term is a penalty factor that aims at keeping the total number of parameters low. In our experiments, as the gestures are quite simple, we compute a set of candidate GMMs with up to 5 states and keep the model with the minimum score (2 components are found for most gestures).

[1] Note that similar requirements also appear when using head-mounted microphone or visual markers, and that the recent development of clothes encapsulating sensors could lower this constraint, see e.g. [16].

[2] People unfamiliar with HMM should refer to [21]

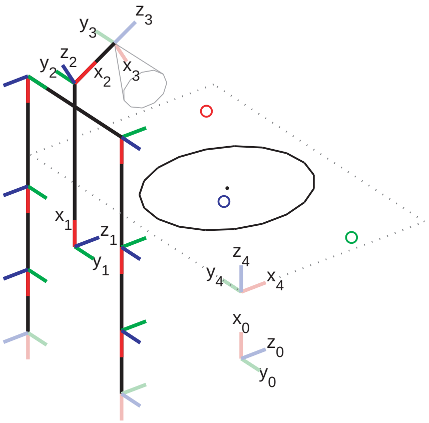

Fig. 4. Extraction of the gaze direction. The table and the 3 objects are represented with dotted lines and circles. The ellipse represents the intersection of the vision cone and the plane defining the table. $O_0x_0y_0z_0$, $O_1x_1y_1z_1$, $O_2x_2y_2z_2$, $O_3x_3y_3z_3$ and $O_4x_4y_4z_4$ are respectively the world, neck, gaze and table frames of reference.

TABLE I

RECOGNITION RESULTS FOR THE TWO TEACHING CONDITIONS. S,E_s,E_i,E_d AND R STAND FOR THE NUMBER OF SAMPLES, SUBSTITUTION, INSERTION, DELETION ERRORS, AND RECOGNITION RATE.

CONDITION A (NATURAL GESTURES)

	S	E_s	E_i	E_d	R
Gazing and pointing	26	0	0	3	88%
Mutual gaze	26	0	3	1	84%
Head nods	26	0	1	0	96%
Head shakes	26	0	1	0	96%
Object selection	26	3	0	0	88%

CONDITION B (KEYBOARD)

	S	E_s	E_i	E_d	R
Gazing and pointing	26	4	0	0	84%
Mutual gaze	26	4	0	0	84%
Head nods	26	0	0	0	100%
Head shakes	26	0	0	0	100%
Object selection	26	4	0	0	84%

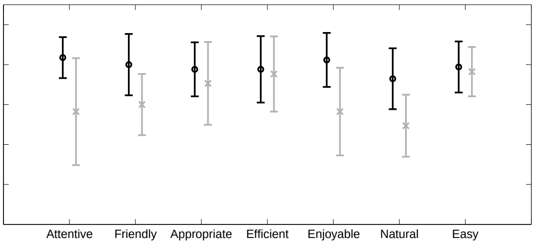

Fig. 5. Results of the questionnaire. For the different attributes, the mean and standard deviation are represented in black for experiment A (natural condition), and in grey for experiment B (keyboard condition).

Similarly to GMM, HMM uses a mixture of Gaussians to describe the distribution of the data, but it also encapsulate the transitions probabilities between the Gaussians. It offers, thus, a way of describing probabilistically the temporal variations of the data. Let $\{\Pi, A, B\}$ be, respectively, the initial state distribution, the transition probabilities between the states and the output distribution. $\{\Pi, A\}$ are computed by *Baum-Welch* algorithm, and $B = \{\mu_k, \Sigma_k\}_{k=1}^{K}$ are the distributions previously found by GMM.

Once trained, the model can recognize gestures by estimating the likelihood that the observed data could have been generated by the model. An absolute threshold and a relative threshold (difference between the first two highest log-likelihoods) are used to determine whether a gesture is recognized or not. The aim of the absolute threshold is to select gestures sharing enough similarities with the model, while the aim of the relative threshold is to select a gesture belonging to a model only if the gesture is sufficiently distant from the other models.

B. Extracting pointing and gazing information

Pointing and gazing directions are modeled as cones with vertex point and directions defined by the hand/gaze frames of reference, see Fig. 4. The intersections of the cones with the table provide information about the object selected by the user, using a probabilistic approach (see the full description of the algorithm in Appendix).

IV. EXPERIMENTAL RESULTS

20 volunteers (mainly students with a mean age of 20) were contacted to test the system. The aims of this preliminary study were: 1) To evaluate the recognition capabilities of the system when faced with untrained user, 2) To compare the efficiency and conviviality of the teaching process presented here with the one used in our previous work, i.e. to see how much perceptive and active communicative behaviors present advantages in a *Programming by Demonstration* scenario. Each person was instructed to wear the motion sensors and to proceed along the teaching scenario depicted in Fig. 2, under two different teaching conditions (order chosen randomly): *A) Natural condition:* The robot recognizes the natural turn-yielding cues, turn-taking cues and yes/no answers of the user, and produces communicative gestures as feedback (both perceptive/active behaviors). *B) Keyboard condition:* The participants use keys to signal turn-yielding cues, turn-taking cues and yes/no answers, with the feedback displayed on a screen (no perceptive/active behavior).

The participants were filmed during the instruction and interaction with the robot. Recognition results using the two teaching conditions are reported in Table I. Of course, the natural teaching condition leads to *insertion/deletion* errors that do not happen when using a keyboard, i.e. gestures are sometimes recognized without the intention of the user (*insertion*) or nothing is recognized when the user produces the gesture (*deletion*). Substitution errors are mainly biased by the production of social cues by the robot to indicate the steps in the scenario. In the natural teaching condition, the robot looks at the table or at the user to signal turn-taking or to request an evaluation, which helps the user produce the required gesture at the right moment during the interaction process, still looking at the table. Substitution errors in condition B are due to this lack of social feedback, i.e. the user can not look simultaneously at the table and at the screen, and loses track of the interaction more easily, i.e. does not produce coherent signals for turn-yielding and joint attention.

After the interaction, questionnaires were given to the participants to rank the interaction process with a 5 point scale (5=very much, 4=somewhat, 3=average, 2=a little, 1=not at all): *1) Did the robot look* ***attentive****? 2) Did the robot look* ***friendly****? 3) Did the robot have* ***appropriate*** *reactions? 4) Did the interaction appear* ***efficient****? 5) Did the interaction appear* ***enjoyable****? 6) Did the interaction appear* ***natural****? 7) Was the teaching process* ***easy****?*

Results of the questionnaire are presented in Fig. 5. For all attributes, the scores for the natural condition are higher. A significant difference (ANOVA, $p < 0.01$, $F_{1,38} > 7.2$) between the two conditions are detected for attributes *attentive*, *friendly*, *enjoyable*, and *natural* (attributes 1,2,5,6). For attributes *appropriate*, *efficient*, and *easy* (attributes 3,4,7), there is no evidence of a difference between the two teaching conditions, which is comforting since the natural teaching condition is less robust, see Table I.

V. Discussion

The questionnaire and the additional remarks noted by the participants showed that adding social cues to our previous *Programming by Demonstration* framework can present advantages, even if the teaching process is less robust and lasts longer than the use of a keyboard. The process did not seem more efficient nor easier, but the social factor increased the mutual attention and enjoyment felt during the interaction with the robot. In the natural teaching condition, a few users had signaled to the robot that it committed a mistake even if it was not the case, in order to play again with the robot. This never happened in teaching condition B, where the use of a keyboard seemed often boring to the user. In condition B, several participants compared the keys actions as a data collecting process (start/stop recording) instead of the social cues depicted by these keys, even if they were not instructed that the robot was collecting data. In the natural teaching condition, the users did not notice that the turn-taking and turn-yielding gestures were aimed at starting/stopping the data collecting process.

In the natural teaching condition, substitution errors for the *gazing and pointing* motion and for the *joint attention* motion were mainly due to the high similarity of head poses. As the user was close to the robot, the joint angles collected when looking at the table and when looking at the robot differed only of a few degrees. Indeed, extracting gaze information by measuring only the orientation of the head is a strong assumption in our system. Head orientation can not be considered directly as a social cue, but it affects gaze following, i.e. the head is naturally turned towards a goal when there is no other constraint.

VI. Conclusion

In this paper, we presented a gesture recognition system using motion sensors, we described a method to extract pointing and gazing information in a probabilistic manner, and we suggested the use of an imitation game to make the teaching process more enjoyable. We showed that incorporating basic social behaviors to our existing *Programming by Demonstration* framework produced life-like behavior which was more enjoyable and intuitive for an untrained user.

Acknowledgments

The work described in this paper was supported in part by the Secretariat d'Etat a l'Education et la Recherche Suisse (SER), under Contract FP6-002020, Integrated Project COGNIRON of the European Commission Division FP6-IST Future and Emerging Technologies, and was supported in part by the Swiss National Science Foundation, through grant 620-066127 of the SNF Professorships program.

Appendix
Algorithms for computing the direction of the user's head and hand

The extraction of the pointing and gazing direction are conducted in a similar way. Here, we describe the process for the gaze information. The gaze is modeled by a cone of vision, defined by vertex point $t_1 = O_0O_3$, direction $d_1 = O_3x_3$ and half-cone angle θ, see Fig. 4. The pointing direction is defined in a similar way by using the hand frame of reference. A point x on the cone satisfy the condition:

$$d_1 \left(\frac{x - t_1}{|x - t_1|} \right) \;=\; cos(\theta)$$

which can be re-written in a matrix form as:

$$\begin{aligned} (x-t_1)^T M (x-t_1) &= 0 \\ \text{with} \;\; M &= d_1 d_1^T - (cos(\theta))^2 I \end{aligned} \tag{1}$$

where I is the identity matrix.

The table is defined by a plane with origin $t_2 = O_0O_4$, first direction $d_{21} = O_4y_4$ and second direction $d_{22} = O_4z_4$, see Fig. 4. A point x on the plane must satisfy the condition:

$$x \;=\; t_2 + x_1\, d_{21} + x_2\, d_{22} \tag{2}$$

By combining (1) and (2), we find the intersection of the cone and the plane, which is defined by:

$$c_1\, x_1^2 + 2c_2\, x_1x_2 + c_3\, x_2^2 + 2c_4\, x_1 + 2c_5\, x_2 + c_6 \;=\; 0$$

with $t_{12} = t_2 - t_1$, $c_1 = d_{21}^T M d_{21}$, $c_2 = d_{21}^T M d_{22}$, $c_3 = d_{22}^T M d_{22}$, $c_4 = t_{12}^T M d_{21}$, $c_5 = t_{12}^T M d_{22}$ and $c_6 = t_{12}^T M t_{12}$.

It defines a quadratic equation representing a conic, that can be re-written in an homogenous matrix form:

$$\begin{aligned} x^T C x &= 0 \quad \text{with} \quad x = (x_1, x_2, 1)^T \\ \text{and} \;\; C &= \begin{pmatrix} c_1 & c_2 & c_4 \\ c_2 & c_3 & c_5 \\ c_4 & c_5 & c_6 \end{pmatrix} \\ &= \begin{pmatrix} C_R & C_t \\ C_t^T & C_\delta \end{pmatrix} \in \begin{pmatrix} \mathbb{R}^{2\times2} & \mathbb{R}^{1\times2} \\ \mathbb{R}^{2\times1} & \mathbb{R}^{1\times1} \end{pmatrix} \end{aligned} \tag{3}$$

The intersection of a cone and a plane can form either an ellipse, parabola or hyperbola. We are interested in elliptical intersection, which happens iff:

$$|C| \neq 0 \quad , \quad \left| \begin{pmatrix} c_1 & c_2 \\ c_2 & c_3 \end{pmatrix} \right| > 0 \quad , \quad \frac{|C|}{c_1 + c_3} < 0$$

In such situation, we determine the canonical form of the conic C_c by transforming the conic matrix C through a rotation R and a translation t, i.e. by applying an Euclidean transformation H:

$$C_c = \begin{pmatrix} C_{c1} & 0 & 0 \\ 0 & C_{c2} & 0 \\ 0 & 0 & C_{c3} \end{pmatrix} = H^T C H$$

$$\text{with} \quad H = \begin{pmatrix} R & t \\ 0^T & 1 \end{pmatrix} \tag{4}$$

C_c defines the canonical conic $C_{c1}x_{c1}^2 + C_{c2}x_{c2}^2 + C_{c3} = 0$, that can be re-written as an ellipse equation:

$$\frac{x_{c1}^2}{a^2} + \frac{x_{c2}^2}{b^2} = 1 \tag{5}$$

$$\text{with} \quad a = \sqrt{-\frac{C_{c3}}{C_{c1}}} \quad , \quad b = \sqrt{-\frac{C_{c3}}{C_{c2}}}$$

To find the homogenous transformation H, the first step is to diagonalize C_R to find the rotation R aligning the conic to the canonical frame. This is achieved by *Principal Component Analysis*, i.e by calculation of the *eigenvalues* of C_R:

$$C_R = R \Lambda R^T \tag{6}$$

The translation t centering the ellipse to the canonical frame is calculated using (3), (4) and (6):

$$t = -R\Lambda^{-1}R^T C_t$$

Using (5), the ellipse in the canonical form can be represented as a covariance matrix:

$$\Sigma_c = \begin{pmatrix} a^2 & 0 \\ 0 & b^2 \end{pmatrix}$$

Finally, the conic with an elliptical form can be represented as a 2D Gaussian distribution $\{\mu, \Sigma\} = \{t, R \;\; \Sigma_c \;\; R^T\}$. With that representation, we define the probabilistic measure of interest (level of saliency) for an object at position x_o by computing the Gaussian density:

$$\mathcal{N}(x_o; \mu, \Sigma) = \frac{1}{\sqrt{(2\pi)^2|\Sigma|}} \; e^{-\frac{1}{2}\left((x_o-\mu)^T \Sigma^{-1} (x_o-\mu)\right)} \tag{7}$$

By taking the logarithm of (7) and meaning over all the collected samples, the object with highest log-likelihood is selected.

References

[1] S. Calinon, F. Guenter, and A. Billard, "On learning, representing and generalizing a task in a humanoid robot," *IEEE Transactions on Systems, Man and Cybernetics, Part B. Special issue on robot learning by observation, demonstration and imitation*, vol. 36, no. 5, 2006, in press.

[2] A. Billard, S. Calinon, and F. Guenter, "Discriminative and adaptive imitation in uni-manual and bi-manual tasks," *Robotics and Autonomous Systems*, vol. 54, no. 5, 2006.

[3] S. Calinon, F. Guenter, and A. Billard, "On learning the statistical representation of a task and generalizing it to various contexts," in *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*, 2006.

[4] S. Schaal, "Is imitation learning the route to humanoid robots?" *Trends in Cognitive Sciences*, vol. 3, pp. 233–242, 1999.

[5] S. Calinon and A. Billard, "Recognition and reproduction of gestures using a probabilistic framework combining PCA, ICA and HMM," in *Proceedings of the International Conference on Machine Learning (ICML)*, 2005.

[6] A. Brooks, J. Gray, G. Hoffman, A. Lockerd, H. Lee, and C. Breazeal, "Robot's play: Interactive games with sociable machines," *ACM Computers in Entertainment*, vol. 2, no. 3, 2004.

[7] C. Breazeal, D. Buchsbaum, J. Gray, D. Gatenby, and B. Blumberg, "Learning from and about others: Towards using imitation to bootstrap the social understanding of others by robots," *Artificial Life*, vol. 11, no. 1-2, 2005.

[8] B. Scassellati, "Imitation and mechanisms of joint attention: A developmental structure for building social skills on a humanoid robot," *Lecture Notes in Computer Science*, vol. 1562, pp. 176–195, 1999.

[9] H. Kozima and H. Yano, "A robot that learns to communicate with human caregivers," in *International Workshop on Epigenetic Robotics*, 2001.

[10] H. Ishiguro, T. Ono, M. Imai, and T. Kanda, "Development of an interactive humanoid robot robovie - an interdisciplinary approach," *Springer Tracts in Advanced Robotics*, vol. 6, pp. 179–192, 2003.

[11] M. Ito and J. Tani, "Joint attention between a humanoid robot and users in imitation game," in *International Conferfence on Development and Learning (ICDL)*, 2004.

[12] V. Hafner and F. Kaplan, "Learning to interpret pointing gestures: experiments with four-legged autonomous robots," in *Biomimetic Neural Learning for Intelligent Robots. Intelligent Systems, Cognitive Robotics, and Neuroscience*, ser. Series: Lecture Notes in Computer Science. Subseries: Lecture Notes in Artificial Intelligence, Vol. 3575, S. Wermter, G. Palm, and M. Elshaw, Eds. Springer Verlag, 2005.

[13] R. Dillmann, "Teaching and learning of robot tasks via observation of human performance," *Robotics and Autonomous Systems*, vol. 47, no. 2-3, pp. 109–116, 2004.

[14] S. Neggers and H. Bekkering, "Ocular gaze is anchored to the target of an ongoing pointing movement," *Neurophysiology*, vol. 83, no. 2, pp. 639–651, February 2000.

[15] S. Calinon, J. Epiney, and A. Billard, "A humanoid robot drawing human portraits," in *IEEE-RAS International Conference on Humanoid Robots*, 2005.

[16] D. D. Rossi, R. Bartalesi, F. Lorussi, A. Tognetti, and G. Zupone, "Body gesture and posture classification by smart clothes," in *IEEE/RAS-EMBS International Conference on Biomedical Robotics and Biomechatronics (BioRob)*, 2006.

[17] S. Fujie, Y. Ejiri, K. Nakajima, Y. Matsusaka, and T. Kobayashi, "A conversation robot using head gesture recognition as para-linguistic information," in *IEEE International Workshop on Robot and Human Interactive Communication (ROMAN)*, 2004, pp. 159–164.

[18] A. Kapoor and R. Picard, "A real-time head nod and shake detector," in *Workshop on Perceptive user interfaces (PUI)*, 2001, pp. 1–5.

[19] A. Meltzoff, "Towards a developmental cognitive science. the implication of cross-modal matching and imitation for the development of representation and memory in infancy," in *The development and Neural Basis of Higher cognitive Functions*, A. Diamond, Ed. Annals of the N. Y. Acad. of Sci., 608, 1996.

[20] K. Nickel and R. Stiefelhagen, "Pointing gesture recognition based on 3d-tracking of face, hands and head orientation," in *international conference on Multimodal interfaces (ICMI)*, 2003, pp. 140–146.

[21] L. Rabiner, "A tutorial on hidden markov models and selected applications in speech recognition," *Proceedings of the IEEE*, vol. 77:2, pp. 257–285, February 1989.

[22] G. Schwarz, "Estimating the dimension of a model," *Annals of Statistics*, vol. 6, pp. 461–464, 1978.