Contents

1 *Combining Dynamic Head Pose-Gaze Mapping with the Robot Conversational State for Attention Recognition in Human-Robot Interactions*, Samira SHEIKHI, Jean-Marc ODOBEZ, [1]

Motivation Gaze is important in discourse regulation. In particular, it is a good indicator of the addressee (important when many people). Modeling HRI.

Approach They do not want to use intrusive devices. Therefore, they want to use computer vision by using head pose estimation as an approximation of the gaze (Kinect and API (Application Program Interface) provide head pose, see also Yu's work). Nevertheless, this may be problematic since same pose could be used for different gaze directions. The paper try to solve this ambiguity by exploring two directions:

- Head pose-VFOA gaze direction association
- Contextual recognition

Previous models work well applied to meetings but suffer when the body is not fixed. And it does not consider the fact that the VFOA depends not only one the head/gaze direction but also on the gaze direction before the shift.

Head pose-VFOA gaze direction association What is the expected head pose of a person who looks at a given VFOA target? That is $p(H_t|F_t)$ or more precisely $\mathbb{E}[H_t|F_t]$. It cannot be considered as a random noise with zero mean. Indeed, the head is not necessary oriented in the direction of the gaze. It depends on the body, head and gaze dynamic.

The Method We assume to have a set of specific visual target $\mathbb{F}$ of interest for the given context (is it dynamic? it seems fixed for an experiment). The set $\mathbb{F}$ contains the different objects plus an element called *other* when the person is not looking at a specific object. The head pose is given by th r.v. $H_t = (H_t^{pan}, H_t^{tilt})$ and the VFOA is a r.v. $F_t \in \mathbb{F}$. The expected head pose, given the gaze direction $\mu_t = (\mu_t^{pan}, \mu_t^{tilt})$, is given by $\mu_t^h = (\mu_t^{h,pan}, \mu_t^{h,tilt})$.

Baseline: HMM with Geometrical Mapping The HMM equations can be written as

$$\left\{ \begin{array}{rcl} p(H_t|F_t = f, \mu_t^h, \Sigma_H) & = & \mathcal{N}(H_t|\mu_t^h(f), \Sigma_H(f)) \\ p(F_t|F_{t-1} = \hat{f}) & = & A_{f\hat{f}} \end{array} \right. \tag{1}$$

The transition matrix $A_{f\hat{f}}$ is defined such that the probability of staying in the same state is large and equal low probability of changing the state. The Gaussian mean of the head pose μ_t^h cannot be learned because we would need annotations data for the VFOA. Therefore, we have to use a gaze model (see [6] and [7] for head-to-gaze model). In this model, we suppose the body orientation to be fixed at R_0. The head pose is a linear combination of the gaze and body reference. That is, for $i \in \{pan, tilt\}$, one have

$$\mu_t^{hb,i}(f) - R_0^i = \alpha^i(\mu_t^i - R_0^i) \Rightarrow \mu_t^{hb,i}(f) = \alpha^i \mu_t^i + (1 - \alpha^i) R_0^i \tag{2}$$

Usually, the head-to-gaze ration $\alpha \in [0.5, 0.7] \times [0.3, 0.5]$.

Model G1: Dynamical Head Reference The baseline method works well for the meeting experiments when the body is static. Nevertheless, in usual HRI the people are moving. Therefore, the assumption of a static body orientation (reference orientation) R_0 is not realistic. The G1 model is considering a dynamical reference direction given by the average of the previous head pose. That is

$$R_t = \frac{1}{W_R} \sum_{i=t-W_R}^{t} H_i. \tag{3}$$

It is considering the fact that we usually align our body with the head pose (it is more comfortable). Then we use the same linear model (2) for the expected head pose. That is, for $i \in \{pan, tilt\}$,

$$\mu_t^{hg1,i}(f) = \alpha^i \mu_t^i + (1 - \alpha^i) R_t^i \tag{4}$$

Model G2: Midline Effect In this model, the expected head pose is more complex. It depends on the previous headpose and the new gaze shift. The midline effect [7] is represented and explained in figure 1.

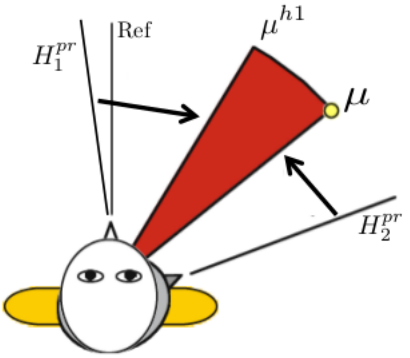


Figure 1: Gaze Model with Midline Effect (Hanes and McCollum, 2006) (pan superscripts are dropped for simplicity). The target direction for the shift is denoted by μ. When the gaze is moved to μ from the initial head pose H_1^{pr}, the head is rotated to μ^{h1} according the geometrical model 2. However, when the gaze shift is centripetal from H_2^{pr} to μ, the head is moved to μ. For initial head positions between μ^{h1} and μ (red zone), an eye-only saccade to μ is made (the head position remains the same).

In order to model this kind of behaviour, we need to introduce the variable H_t^{pr} which is the head pose prior to a gaze shift. One have

$$H_t^{pr} = \frac{1}{W_p} \sum_{i=t-W_R-\Delta^p}^{t-\Delta^p} H_i. \tag{5}$$

This model is only applied to the pan angle (where we have omitted the superscript). Remark that the expression for H_t^{pr} is very similar to the definition of R_t; the difference is a shift in the averaging (more or less, for t, $\exists s < t$ such that $H_t^{pr} = R_s$). Finally, the head pose expected value, for $\mu_t > 0$, is defined through (only for pan angle)

$$\mu_t^{hg2} = \begin{cases} \mu_t^{hg1} & \text{if } H_t^{pr} < \mu_t^{hg1}, \\ \min(\mu_t,\, \alpha_H H_t^{pr} + (1 - \alpha_H)\mu_t^{hg1}) & \text{otherwise.} \end{cases} \tag{6}$$

Remark that if $alpha_H = 0$, it becomes the G1 model and if $alpha_H = 1$, the model is exactly the one defined in [7].

Model G3: Implementing Gaze Shifts The previous model has one drawback: at each time step a gaze shift is assumed. Therefore, the head pose in always updated is position when looking at a target. That is the expected head pose would converge to the gaze direction (? note sure ?). Therefore, the idea is do consider the previous VFOA, $\hat{f} = F_{t-1}$, and define the following expected head pose

$$\mu_t^h g3 = \alpha_1 \mu_t(f) + \alpha_2 \mu_t(\hat{f}) + (1 - \alpha_1 - \alpha_2) R_t. \tag{7}$$

Thus, in absence of gaze shift, the models is equivalent to the G1 model. If $f \neq \hat{f}$, then the head would be closer to direction of the previous VFOA that what predicted by G1 model. The graphical model for the models G2 and G3 is presented in figure 2.

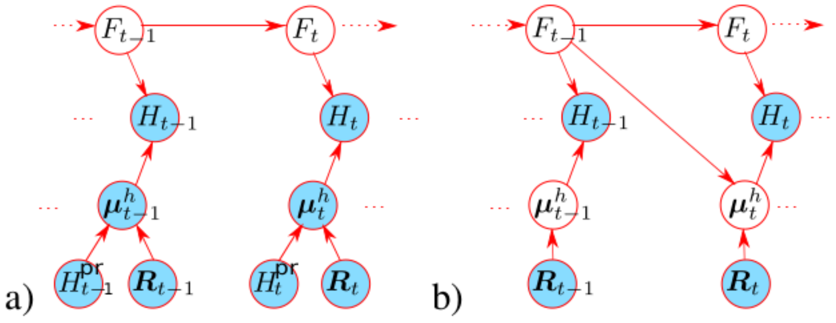


Figure 2: a) Model G2. b) Model G3.

Context Modeling The overall conversational context $C_t = (s_t, a_t, o_t)$ is defined by the speaking context $s_t \in \{0, 1\}$ as whether the robot is speaking or not, the addressee context $a_t \in AC = \{pers_1, pers_2, group\}$. This context is automatically derived from the dialog system. The dialog system is aware of who is addressed and how to address him (looking at the person or calling the name, or looking at average of the persons when talking to the group). People who are addressed may look more often the robot. Finally, the topic context $o_t \in OC = \{pai_1, pai_2, pai_3, paintings, non\}$ corresponds to whether the robot informs or refer to a specific painting.

Overall Model The overall model is given by the IOHMM (Input Output HMM) graphical model of figure 3.

The posterior for the graphical model is defined through

$$\begin{array}{rcl} p(F_{1:T}|H_{1:T}, C_{1:T}, \mu^h_{1:T}, R_{1:T}) & \propto & \prod_{t=1}^{T} p(H_t|F_t, \mu_t^h) p(F_t|F_{t-1}, C_t) \\ p(H_t|F_t = f, \mu_t^h, \Sigma_H) & = & \mathcal{N}(H_t|\mu_t^h(f), \Sigma_H(f)) \\ p(F_t|F_{t-1}, C_t) & \propto & p(F_t|F_{t-1}) p(F_t|C_t) = A_{f\hat{f}} \cdot B_{cf} \end{array} \tag{8}$$

In the last equation of (8), we assume F_{t-1} and C_t to be independent.The probability table $B_{cf} = p(F_t = f|C_t = c)$ denotes the robot context prior on what people are looking at depending on the context. The context table B is learned from data annotations (with smoothing to hadle the lack of data for some context (? not clear for me ?)). To avoid overfitting, parameter tying is applied (? not clear for me ?). Let $D_c = \{f_i\}$ be the VFOA data observed under the context c. Then, using a Maximum A Posteriori approach with

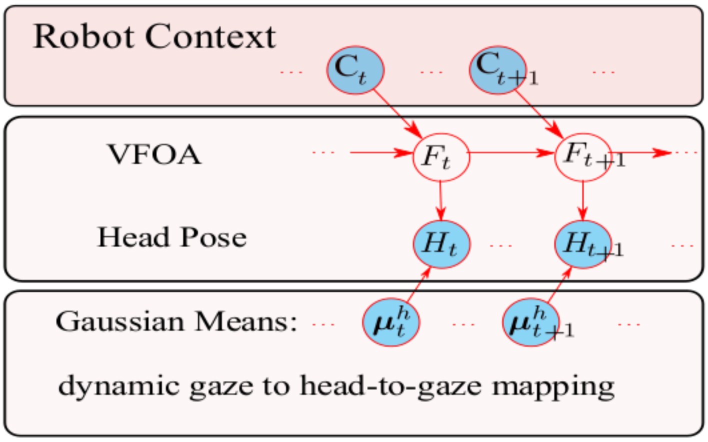


Figure 3: VFOA recogniation from head pose.

a conjugate Dirichlet prior (i.e. maximizing $p(B_c|D_c) \propto p(D_c|B_c)Dir(B_c|\alpha)$), the table entries are defined as $B_{cf} \propto n_f + \alpha_f$, where n_f is the denotes the number of occurrences of the focus f in D_c and $\alpha = 0.1N_f/(K \times N_C)$, where N_f is the number of observation in the whole training set, K is the number of VFOA targets and N_C is the number of contexts. In other words, the prior corresponded to the addition of virtual observations equally spread amongst table entries and amounting to 10% of the total number of observations (? not clear for me ?).

Results In the first step, we ignore the contextual part (first equation in (8)). The head pose is estimated with Vicon head pose and with head pose tracker data (see [8]). The reference is the head of the robot (since the robot is moving, the reference is not fixed). The problem with video tracking of the head pose was that the robot is moving and therefore, sometimes, it does not see the people (i.e head pose is missing) and, since the image is perturbed by nodding, may be less accurate than Vicon tracker. The average head pose error of the head tracker is shown in figure 4. For the gaze direction (that is the location of the differend persons and object), is obtained with Vicon sensors. In a more general application, the robot should do the tracking.

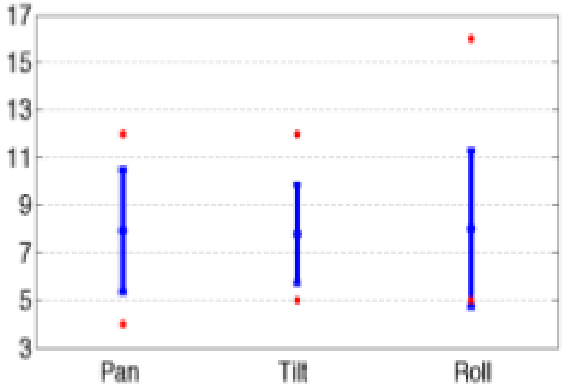
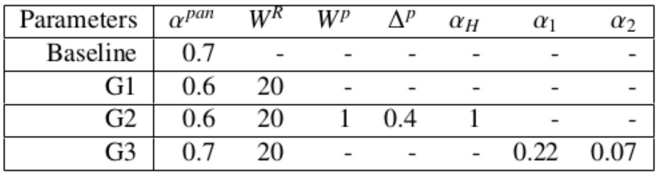


Figure 4: Average head pose error when using head tracker [8].

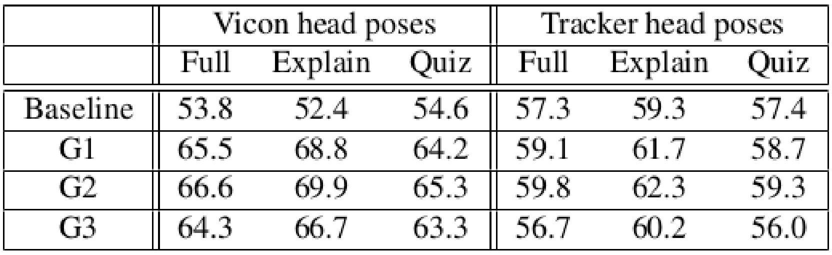
The parameter setting is given in figure 5. In majority, it is obtained from cross-validation.



Parameters	α^{pan}	W^R	W^p	Δ^p	α_H	α_1	α_2
Baseline	0.7	-	-	-	-	-	-
G1	0.6	20	-	-	-	-	-
G2	0.6	20	1	0.4	1	-	-
G3	0.7	20	-	-	-	0.22	0.07

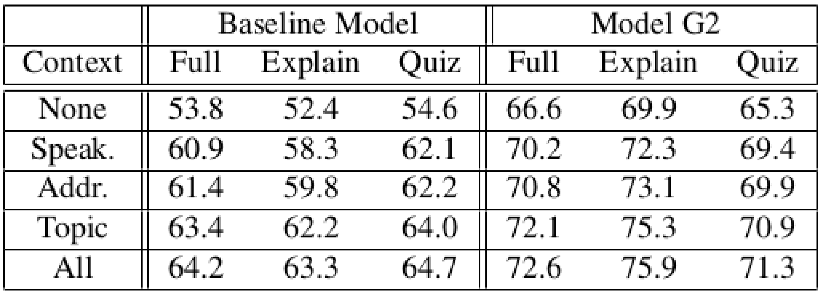
Figure 5: Parameter setting for experiments.

The results, when ignoring the context, is shown in figure 6. The results with the context is shown in figure 7 when using the Vicon head pose and in figure 8. The comparaison is done by considering only one context (speaking, addressee or context topic) and also for all cues together.



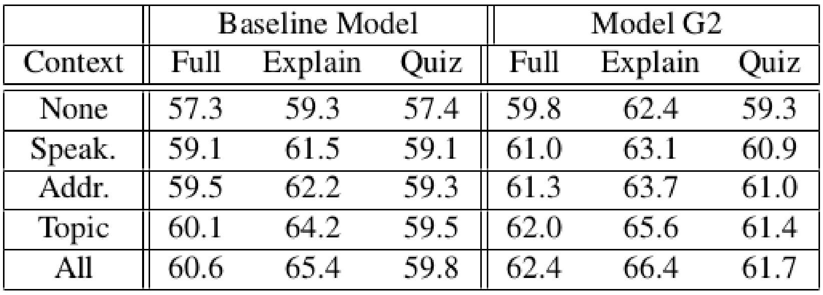
	Vicon head poses			Tracker head poses		
	Full	Explain	Quiz	Full	Explain	Quiz
Baseline	53.8	52.4	54.6	57.3	59.3	57.4
G1	65.5	68.8	64.2	59.1	61.7	58.7
G2	66.6	69.9	65.3	59.8	62.3	59.3
G3	64.3	66.7	63.3	56.7	60.2	56.0

Figure 6: Recognition rates of head-gaze mappings without context.



	Baseline Model			Model G2		
Context	Full	Explain	Quiz	Full	Explain	Quiz
None	53.8	52.4	54.6	66.6	69.9	65.3
Speak.	60.9	58.3	62.1	70.2	72.3	69.4
Addr.	61.4	59.8	62.2	70.8	73.1	69.9
Topic	63.4	62.2	64.0	72.1	75.3	70.9
All	64.2	63.3	64.7	72.6	75.9	71.3

Figure 7: Recognition rates of head-gaze mappings with context using Vicon head pose.



	Baseline Model			Model G2		
Context	Full	Explain	Quiz	Full	Explain	Quiz
None	57.3	59.3	57.4	59.8	62.4	59.3
Speak.	59.1	61.5	59.1	61.0	63.1	60.9
Addr.	59.5	62.2	59.3	61.3	63.7	61.0
Topic	60.1	64.2	59.5	62.0	65.6	61.4
All	60.6	65.4	59.8	62.4	66.4	61.7

Figure 8: Recognition rates of head-gaze mappings with context using the head tracker.

Remark

- Use RGB-D camera for a better approximation of head pose (cf. Yu's works)
- Using gazing direction (Keneth's work). It would provide priors on the gaze. That is (note sure about that)

$$p(F_{1:T}|H_{1:T}, G_{1:T}, C_{1:T}, \mu^h_{1:T}, R_{1:T}) \propto \prod_{t=1}^{T} p(H_t|F_t, \mu^h_t)p(G_t|F_t, \mu_t)p(F_t|F_{t-1}, C_t) \quad (9)$$

- The context can be improve by considering the timing (how long is the dialog act active?) and robot gesture.

2 *Tracking Gaze and Visual Focus of Attention of People Involved in Social Interaction*, Benoit Masse, Sileye Ba, Radu Horaud, [2]

Motivation Computational analysis of social interactions. In particular, the VFOA is important because it is one of the prominent social cues.

Remark It is said that method based on eye tracking are ineffective because data are missing (face away from camera). But we can think about a model that perform better if eyes are visible and will still perform well if data is missing.

Introduction They propose a Bayesian switching dynamic model. The method can be viewed as a computational model of [9] and [10] (check those articles). They provide a tractable learning algorithm. Softwares packages, examples of results and datset for VERNISSAGE and LAEO (looking at each other) are available at `https://team.inria.fr/perception/research/eye-gaze/`. The proposed model assume that gaze shifts are produced by head movements that occurs simultaneously with eye movement (? not clear for me ?).

Method The VFOA of person i at time t is given by $V^i_t \in \mathbb{F}$, where $\mathbb{F}$ is the set of objects for focus of attention. It is split in three part: active ($1 \leq V^i_t \leq N$), passive ($N+1 \leq V^i_t \leq M+N$), none ($V^i_t = 0$)). The set of all VFOA (for everybody present) is given by $V_t = (V^1_t, ..., V^N_t)$. The head poses are given by $H_t = (H^1_t, ..., H^N_t)$ and the gaze directions by $G_t = (G^1_t, ..., G^N_t)$. More over the head reference orientation is given by $R_t = (R^1_t, ..., R^N_t)$. The coordinates are pan and tilt angles (roll does not matter for gazing but it could be interesting to see if there is a meaning for attention). Finally, the location of an object $i \in \{1, ..., N+M\}$ is given by X^i_t and the relative direction of an object j with respect to a person i (unitary vector pointing from person i to the object j) is given by $X^{ij}_t = (X^j_t - X^i_t)/\|X^j_t - X^i_t\|$. The goal is to solve the maximum a posteriori (MAP) problem

$$\hat{V}_t, \hat{G}_t = \underset{V_t, G_t}{\operatorname{argmax}}\, p(V_t, G_t|H_{1:t}, X_{1:t}). \quad (10)$$

The generative model is based on [11]. The same model is used in [1]. It is given by

$$p(H^i_t|G^i_t, R^i_t; \alpha, \Sigma_H) = \mathcal{N}(H^i_t; \mu^i_{H_t}, \Sigma_H) \quad (11)$$

with

$$\mu^i_{H_t} = \alpha G^i_t + (I_2 - \alpha)R^i_t. \quad (12)$$

The next steps will be the definition of the stochastic processes G_t and R_t.For the gaze dynamics the following model is proposed

$$\begin{aligned} p(G_t^i|G_{t-1}^i, \dot{G}_{t-1}^i, V_t^i = j, X_t^i) &= \mathcal{N}(G_t^i; \mu_{G_t}^{ij}, \Gamma_G) \\ p(\dot{G}_t^i|\dot{G}_{t-1}^i) &= \mathcal{N}(\dot{G}_t^i; \dot{G}_{t-1}^i, \Gamma_{\dot{G}}) \end{aligned} \tag{13}$$

with

$$\mu_{G_t}^{ij} = \begin{cases} G_{t-1}^i + \dot{G}_{t-1}^i dt & \text{if } j = 0 \text{ (random walk)}, \\ \beta G_{t-1}^i + (I_2 - \beta) X_t^{ij} + \dot{G}_{t-1}^i & \text{otherwise.} \end{cases} \tag{14}$$

In a similar way, they define the reference orientation dynamics as

$$\begin{aligned} p(R_t^i|R_{t-1}^i, \dot{R}_{t-1}^i) &= \mathcal{N}(R_t^i; \mu_{R_t}^i, \Gamma_R) \\ p(\dot{R}_t^i|\dot{R}_{t-1}^i) &= \mathcal{N}(\dot{R}_t^i; \dot{R}_{t-1}^i, \Gamma_{\dot{R}}) \end{aligned} \tag{15}$$

with

$$\mu_{R_t}^i = R_{t-1}^i + \dot{R}_{t-1}^i dt \text{ random walk).} \tag{16}$$

For the overall model, we assume that the random variable are independent between the participant. Therefore, the overall probabilities are simply the product of person specific probabilities. Concerning the VFOA random variable V_t, it assume to be a first order Markov process. Supposing independence between participant (active object), we get

$$p(V_t|V_{1:t-1}) = p(V_t|V_{t-1}) = \prod_{i=1}^{N} p(V_t^i|V_{t-1}) \tag{17}$$

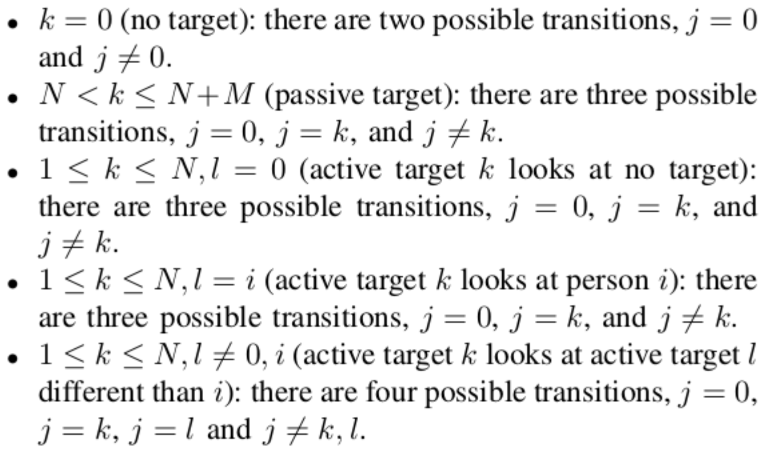
Notice that the matrix $p(V_t|V_{t-1})$ is of size $(N+M)^N \times (N+M)^N$ which is a large number of parameters to predict. But, for example, by assuming independence between the participants we have that $p(V_t^i|V_{t-1})$ is of size $(N+M) \times (N+M)^N$. Additional assumption is maid for example

- If $V_t^i = k \in \{N+1, ..., N+M\} \cup \{0\}$ is a passive target, then V_t^i depends only on V_{t-1}^i. That is to say that the VFOA if we are looking at a passive object only depends on our previous VFOA but is independent of the other participant. I am not sure about this assumption. For example, if three on four participants are looking at a target then the fourth participant may be induced to look at the target as well.
- If $V_t^i = k \in \{1, ..., N\} \backslash i$, then V_t^i will depend on V_{t-1}^i and V_{t-1}^k as well. That is to say that if participant i is looking at participant k, then there VFOA is correlated.

Finally, they use parameter "sharing" (or tying, not sure about the right term) for transition probability. That is that the transition to a passive object is the same for all passive object, for example. In this way they reduce the number of parameters to 15 transitions for the probability $p(V_t^i = j|V_{t-1})$. The assumptions are presented in figure 9.

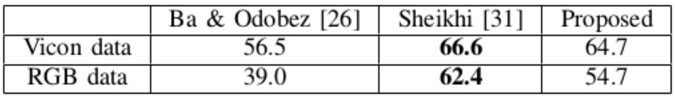
Inference Check the article. It would not be efficient and practical to copy it.

Results The results with VERNISSAGE dataset is given in figure 10.



- $k = 0$ (no target): there are two possible transitions, $j = 0$ and $j \neq 0$.
- $N < k \leq N+M$ (passive target): there are three possible transitions, $j = 0$, $j = k$, and $j \neq k$.
- $1 \leq k \leq N, l = 0$ (active target k looks at no target): there are three possible transitions, $j = 0$, $j = k$, and $j \neq k$.
- $1 \leq k \leq N, l = i$ (active target k looks at person i): there are three possible transitions, $j = 0$, $j = k$, and $j \neq k$.
- $1 \leq k \leq N, l \neq 0, i$ (active target k looks at active target l different than i): there are four possible transitions, $j = 0$, $j = k$, $j = l$ and $j \neq k, l$.

Figure 9: Assumption for the reduction of the transition matrix for $p(V_t^i = j|V_{t-1})$.



	Ba & Odobez [26]	Sheikhi [31]	Proposed
Vicon data	56.5	**66.6**	64.7
RGB data	39.0	**62.4**	54.7

Figure 10: FRR score on the vernissage. The results are compare with [11] (Ba and Odobez) and [1] (Sheikhi). There is some mistake in the reporting of the data from [1]. The value without context are

3 Supervised Gaze Bias Correction for Gaze Coding in Interactions, Remy SIEGFRIED, Jean-marc ODOBEZ, 2017

Motivation Understanding the role of gaze in HHI and HRI. In particular when people are looking at each other.

Method Thank to Keneth's code, we can extract from RGB-D images the head pose $p = (R, T)$ and the gaze angles $g = (\phi, \theta)$. From the gaze angles, one can define the gaze vector $v = v(g)$ (or $v = v(p, g)$?). For a given visual target u_t (object detection), we define the angular error

$$e_t = \arccos(u_t \cdot v) \tag{18}$$

Since we are never looking exactly at a point (we look the face and move frome eyes to mouth, for example), we use a threshold to identify gazing. That is if $e_t < \tau$ then we can assume the person is looking at target t. This method perform well on the dataset of [12]. Nevertheless, it does not perform well with the current dataset. Might be

- The eyes position is estimated from the theoretical eye position in the 3DMM mesh.
- The training set for gaze appearance is not representing well person-specific features.

In order to remedy to this issue, they propose a bias correction. That is they compute on n frames, the average $b_t = \sum g_i/n$, which represent the bias when looking at target t. Then the error is defined by

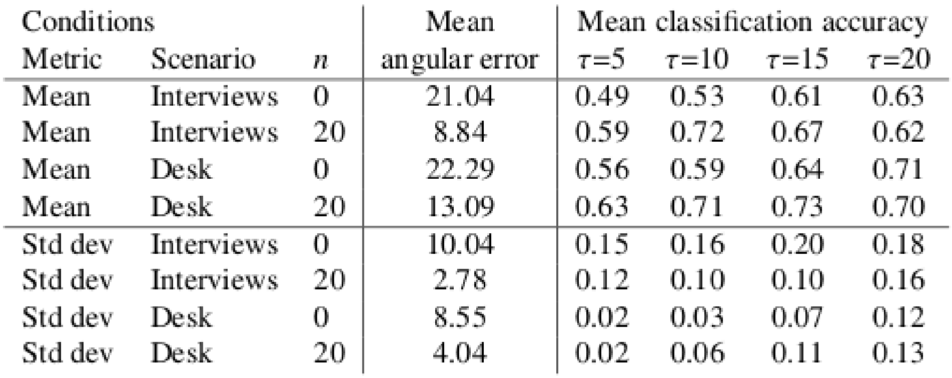
$$e = \arccos(u_t \cdot v(g - b)) \tag{19}$$

Experiment Two set of data:

- Interview: only to persons, face to face.
- Desk: A receptionist answer to client. People are standing and therefore more free to move. More challenging data set.

They have annotation for 2500 frames per videos (8+4=12 videos). If people are looking at each other.

Results The results are shown in figure 11.



Conditions			Mean	Mean classification accuracy			
Metric	Scenario	n	angular error	τ=5	τ=10	τ=15	τ=20
Mean	Interviews	0	21.04	0.49	0.53	0.61	0.63
Mean	Interviews	20	8.84	0.59	0.72	0.67	0.62
Mean	Desk	0	22.29	0.56	0.59	0.64	0.71
Mean	Desk	20	13.09	0.63	0.71	0.73	0.70
Std dev	Interviews	0	10.04	0.15	0.16	0.20	0.18
Std dev	Interviews	20	2.78	0.12	0.10	0.10	0.16
Std dev	Desk	0	8.55	0.02	0.03	0.07	0.12
Std dev	Desk	20	4.04	0.02	0.06	0.11	0.13

Figure 11: Mean angular error and classification accuracy

4 *Towards the Use of Social Interaction Conventions as Prior for Gaze Model Adaptation*, Remy SIEGFRIED, Yu YU, Jean-marc ODOBEZ, [3]

Motivation Extract gaze direction from RGB-D camera recordings.

Method The method used is similar to the one presented in [12]. The head pose is obtained with the use of a 3D morphable model (3DMM) mesh of the depth data. Then a variant of the iterative closest point (ICP) and visual processing based on [13] is used. Then, based on the head pose $p = (R, T)$, a frontal face view is generated using the depth data (but maybe we can do directly the transform from the 2D image?). Then a cropping is done for the eyes region. The baseline method use landmarks on the 3DMM mesh. The current method use landmarks obtained with the Dlib library (check this library). The gaze estimation is obtained following [12]. That is with the use of support vector machine (SVM) applied to a Histogram-of-Gradients(HoG) (edge detection). Finally the attention decision is modeled by the method presented in Remy Siegfried paper (cf. *Supervised Gaze Bias Correction for Gaze Coding in Interactions* above). The bias correction is done in two way. First with annotations on the first minutes of each video (supervised). Secondly, by using the fact that people may look at the speaker more often. Therefore using some frame where someone is speaking and supposing that the participant is looking at him. The bias is computed using a mean on some frames or a least median of squares (LMoS).

Results the results are presented in figure 12. The key words mean

- Lm and AvgLm: Dlib routines. AvgLm use an average over time for landmarks detection.
- GT: use of the annotations as ground truth for frame selection (bias correction).
- Spk: use of the speaker concept as ground truth for frame selection (bias correction)
- Mean and Med: bias correction estimation.

Remark Maybe the bias comes from the cropping. Indeed, if the cropping is not exactly the same between people, the SVM should be invariant to some transformations. Is it the case (I should check).

	Interviews		Desk		Overall	
Method	error	accuracy	error	accuracy	error	accuracy
Baseline	21.04	0.53	26.19	0.59	23.62	0.56
Lm	10.41	0.72	15.24	0.62	12.82	0.67
AvgLm	9.79	0.67	13.72	0.66	11.76	0.67
GT-Mean	7.42	0.75	10.31	0.75	8.86	0.75
GT-Med	7.53	0.75	10.45	0.74	8.99	0.74
Spk-Mean	8.49	0.66	10.59	0.74	9.54	0.70
Spk-Med	9.58	0.68	10.25	0.75	9.91	0.72
Lm + GT-Mean	5.61	0.85	8.85	0.82	7.23	0.84
Lm + GT-Med	5.67	0.84	8.94	0.82	7.30	0.83
Lm + Spk-Mean	7.92	0.73	9.26	0.80	8.59	0.77
Lm + Spk-Med	6.25	0.82	9.03	0.81	7.64	0.82
AvgLm + GT-Mean	6.08	0.82	9.34	0.80	7.71	0.81
AvgLm + GT-Med	6.44	0.82	10.26	0.78	8.35	0.80
AvgLm + Spk-Mean	8.39	0.72	10.65	0.74	9.45	0.73
AvgLm + Spk-Med	6.66	0.82	9.49	0.79	8.07	0.80

Figure 12: The error is the angle in degree. The accuracy is the FRR score (I think).

5 *A Flexible New Technique for Camera Calibration*, Zhengyou ZHANG, [4]

Reference See also the book [14, Chapters 6 and 7].

Introduction There are two types of calibration.

- Pinhole calibration
  - Extrinsic calibration
  - Intrinsic calibration
- Distortion calibration
  - Radial distortion calibration
  - Tangential distortion calibration

Pinhole calibration We assume to have a set of correspondences points $X_i \leftrightarrow x_i$ where $X_i \in \mathbb{P}^3$ and $x_i \in \mathbb{P}^2$ are respectively the homogeneous coordinate of 3D real world points and 2D image points. The *camera matrix* is a homography such that $x_i \propto PX_i$ (there is no equality, only up to a constant). The camera matrix is given through the extrinsic parameter R and t denoting the rotation an translation of the camera with respect to the Real World Coordinate (RWC) and the intrinsic parameter A denoting the intrinsic components of the camera, namely the focal length, principal points (optimal center) and the skew coefficient (usually zero). The camera matrix is obtained from

$$P = A[R \,|\, t] \tag{20}$$

where camera intrinsic matrix is given by

$$A = \begin{bmatrix} \alpha & \gamma & u_0 \\ 0 & \beta & v_0 \\ 0 & 0 & 1 \end{bmatrix} \tag{21}$$

with (u_0, v_0) are the coordinates of the principal point (in pixel), α and β are the scales factors of the image in direction u and v. Finally, the parameter γ represents the skewness of the two image axes ($\gamma = \beta \tan(\theta)$) Without loss of generality, we assume the model plane is on $Z = 0$ and we write $R = [r_1 \,|\, r_2 \,|\, r_3]$. Hence, for an arbitrary constant s, one have

$$s\, x_i = \begin{bmatrix} u_i \\ v_i \\ 1 \end{bmatrix} = A[r_1 \,|\, r_2 \,|\, r_3 \,|\, t] \begin{bmatrix} X \\ Y \\ 0 \\ 1 \end{bmatrix} = A[r_1 \,|\, r_2 \,|\, t] \begin{bmatrix} X \\ Y \\ 1 \end{bmatrix} \tag{22}$$

Therefore the problem is to find homography $H = A[r_1 \,|\, r_2 \,|\, t]$. For $H = [h_1 \,|\, h_2 \,|\, h_3]$, one have $h_i = sAr_i \Rightarrow A^{-1}h_i = r_i$ and since the rotation matrix is orthonormal, one have

$$\begin{aligned} h_1^T A^{-T} A^{-1} h_2 &= 0, \\ h_1^T A^{-T} A^{-1} h_1 &= h_2^T A^{-T} A^{-1} h_2. \end{aligned} \tag{23}$$

Closed form solution This is actually not used in real calibration. Indeed, due to noise measurement the system has not always a solution. In particular, if we have many samples the problem will certainly not have a solution due to the noise. Therefore one should use a residual minimization method. Let

$$B = A^{-T}A^{-1} = \begin{bmatrix} B_{11} & B_{12} & B_{13} \\ B_{21} & B_{22} & B_{23} \\ B_{31} & B_{32} & B_{33} \end{bmatrix} = \begin{bmatrix} B_{11} & B_{12} & B_{13} \\ B_{12} & B_{22} & B_{23} \\ B_{13} & B_{23} & B_{33} \end{bmatrix} \tag{24}$$

$$\begin{bmatrix} \frac{1}{\alpha^2} & -\frac{\gamma}{\alpha^2\beta} & \frac{v_0\gamma-u_0\beta}{\alpha^2\beta} \\ -\frac{\gamma}{\alpha^2\beta} & \frac{\gamma^2}{\alpha^2\beta^2}+\frac{1}{\beta^2} & -\frac{\gamma(v_0\gamma-u_0\beta)}{\alpha^2\beta^2}-\frac{v_0}{\beta^2} \\ \frac{v_0\gamma-u_0\beta}{\alpha^2\beta} & -\frac{\gamma(v_0\gamma-u_0\beta)}{\alpha^2\beta^2}-\frac{v_0}{\beta^2} & \frac{(v_0\gamma-u_0\beta)^2}{\alpha^2\beta^2}+\frac{v_0^2}{\beta^2}+1 \end{bmatrix}$$

Figure 13: Components of the symmetric matrix B.

The matrix components B_{ij} can be computed explicitly and the values are given in figure 13 (see [4]) and the matrix B is symmetric. Recall that $H = [h_1 \,|\, h_2 \,|\, h_3]$ with $h_i = [h_{i1}, h_{i2}, h_{i3}]^T$. By defining

$$b = [B_{11}, B_{12}, B_{22}, B_{13}, B_{23}, B_{33}]^T \tag{25}$$

and

$$\begin{aligned} v_{ij} = &[h_{i1}h_{j1},\ h_{i1}h_{j2} + h_{i2}h_{j1},\ h_{i2}h_{j2}, \\ &h_{i3}h_{j1} + h_{i1}h_{j3},\ h_{i3}h_{j2} + h_{i2}h_{j3},\ h_{i3}h_{j3}]^T \end{aligned}$$

one have

$$h_i^T B h_j = v_{ij}^T b. \tag{26}$$

Therefore, the system of equation in (23) as an homogeneous equation in b:

$$\begin{bmatrix} v_{12}^T \\ (v_{11} - v_{22})^T \end{bmatrix} b = 0. \tag{27}$$

For n images, we simply stack the the equations in V such that

$$Vb = 0. \tag{28}$$

- If $n \geq 3$: in general a unique solution b (that is not really the case due the uncertainty).
- If $n = 2$: We can impose the skewness $\gamma = 0$ (e.i $B_{12} = 0$).
- If $n = 1$: We can only solve two camera intrinsic parameters. We cab suppose the skewness $\gamma = 0$ and the principal point (u_0, v_0) to be known.

In order to avoid trivial solution ($b = 0$), we need to add a constraint. Since the solution is obtained up to an arbitrary constant, we can impose $||b|| = 0$. That is the system, we want to solve is

$$\begin{aligned} &\text{minimize} && ||Vb|| \\ &\text{subject to} && ||b|| = 1. \end{aligned} \tag{29}$$

This problem is equivalent to

$$\text{minimize } \frac{\|Vb\|}{\|b\|} \tag{30}$$

and the solution of the above problem is given by the eigenvector with the least eigenvalue of the matrix $V^T V$. Once the matrix B (or vector b) is computed. One can compute the extrinsic and intrinsic parameters with the formulas of figure 14.

$$\begin{aligned}
\mathbf{r}_1 &= \lambda \mathbf{A}^{-1}\mathbf{h}_1 \\
\mathbf{r}_2 &= \lambda \mathbf{A}^{-1}\mathbf{h}_2 \\
\mathbf{r}_3 &= \mathbf{r}_1 \times \mathbf{r}_2 \\
\mathbf{t} &= \lambda \mathbf{A}^{-1}\mathbf{h}_3
\end{aligned}
\qquad
\begin{aligned}
v_0 &= (B_{12}B_{13} - B_{11}B_{23})/(B_{11}B_{22} - B_{12}^2) \\
\lambda &= B_{33} - [B_{13}^2 + v_0(B_{12}B_{13} - B_{11}B_{23})]/B_{11} \\
\alpha &= \sqrt{\lambda/B_{11}} \\
\beta &= \sqrt{\lambda B_{11}/(B_{11}B_{22} - B_{12}^2)} \\
\gamma &= -B_{12}\alpha^2\beta/\lambda \\
u_0 &= \gamma v_0/\beta - B_{13}\alpha^2/\lambda \ .
\end{aligned}$$

Figure 14: Computation of extrinsic and intrinsic parameters once the matrix B is obtained. I suppose that the matrix A is obtained using a Cholesky decomposition of B. No! since A contains the intrinsic parameters already given.

Maximum likelihood estimation As explain before, we need to solve the problem as minimization one since the system will certainly be unfeasible (except for trivial solution). Shortly, for n images with m elements in each of them, the goal is to minimize

$$\sum_{i=1}^{n}\sum_{j=1}^{m} \|x_{ij} - H_i X_{ij}\|^2. \tag{31}$$

This can be done, for example, with use of Levenberg-Marquart algorithm (`Minpack`). The initial guests are obtained with the close form solution defined in the previous section.

Radial distortion Let (x, y) and $(\hat{x}, \hat{y})$ be the ideal (distortion-free) and real (distortion) normalized image coordinates. We have

$$\left\{ \begin{array}{rcl} \hat{x} &=& x\left(1 + k_1(x^2+y^2) + k_2(x^2+y^2)^2\right), \\ \hat{y} &=& y\left(1 + k_1(x^2+y^2) + k_2(x^2+y^2)^2\right). \end{array} \right. \tag{32}$$

One can consider more terms in the distortion ($+k_3(x^2+y^2)^3$, it is a Taylor expansion). Let (u, v) be the ideal (non-observable distortion-free) pixel image coordinates and $(\hat{u}, \hat{v})$ be the corresponding real (distortion) observed image coordinate. The center of the radial distortion is given by the principal point (u_0, v_0). Therefor, one have $\hat{u} = u_0 + \alpha\hat{x} + \gamma\hat{y}$ and $\hat{v} = v_0 + \beta\hat{y}$. Hence, substitution in the above equations leads

$$\left\{ \begin{array}{rcl} \hat{u} &=& u + (u-u_0)\left(k_1(x^2+y^2) + k_2(x^2+y^2)^2\right), \\ \hat{v} &=& v + (v-v_0)\left(k_1(x^2+y^2) + k_2(x^2+y^2)^2\right). \end{array} \right. \tag{33}$$

The above equation can be reduce to a linear equation in k_1 and k_2. Indeed, one have

$$\left[\begin{array}{cc} (u-u_0)(x^2+y^2) & (u-u_0)(x^2+y^2)^2 \\ (v-v_0)(x^2+y^2) & (v-v_0)(x^2+y^2)^2 \end{array} \right] \left[\begin{array}{c} k_1 \\ k_2 \end{array} \right] = \left[\begin{array}{c} \hat{u}-u \\ \hat{v}-v \end{array} \right] \tag{34}$$

Given m points in n images, we can stack the above system and obtain in total $2mn$ equations. That is to say we have $Dk = d$ where $k = (k_1, k_2)^T$, D is the stacked equations and d the stacked right-hand-side of the equation. The linear least-square solution is given by

$$k = (D^T D)^{-1} D^T d \tag{35}$$

One possibility is to estimate the radial distortion by alternation. That is to say to solve the pinhole calibration using the above transformation (33).

Complete Maximum Likelihood Estimation The alternation method converge too slowly. The best way of solving the problem is to solve it in once using Levenberg-Marquardt algorithm. We define, for image i and point j

$$\hat{m}(A, k_1, k_2, R_i, t_i, X_j)_u = (HX_j)_u + ((HX_j)_u - u_0)\left(k_1(x^2 + y^2) + k_2(x^2 + y^2)^2\right) \tag{36}$$

and the goal is to minimize

$$\sum_{i=1}^{n} \sum_{j=1}^{m} \|x_{ij} - \hat{m}(A, k_1, k_2, R_i, t_i, X_j)\|^2. \tag{37}$$

Remark: Degenerate configuration If the model plane at the second position is parallel to its first position, then the second homography does not provide additional constraints.

6 *Joint Depth and Color Camera Calibration with Distortion Correction*, Daniel HERRERA, Juho KANNALA, Janne HEIKKILA, [5]

Introduction The article describe a method for RGB-D camera calibration. It takes into account radial and tangential distortion corrections. It is the commonly used RGB-D camera calibration.

Color Camera Calibration (intrinsic) The model is similar to the one presented in [15]. Let $\bar{x}_c = (x_c, y_c, z_c)$ be the RWC of a points and $\bar{p}_c = (u_c, v_c)$ its image coordinates (in pixel). The goal is to find $\bar{p}_c$ with respect to $\bar{x}_c$. Firs step consists in normalization

$$\bar{x}_n = \left(\frac{x_c}{z_c}, \frac{y_c}{z_c}\right). \tag{38}$$

Then, the tangential distortion is modeled by (where $r^2 = x_n^2 + y_n^2$)

$$\bar{x}_g = \begin{bmatrix} 2k_3 x_n y_n + k_4(r^2 + 2x_n^2) \\ k_3(r^2 + 2y_n^2) + k_4 x_n y_n \end{bmatrix} \tag{39}$$

and the radial distortion is given by

$$\bar{x}_r = (1 + k_1 r^2 + k_2 r^4 + k_5 r^6)\bar{x}_n. \tag{40}$$

Therefore, the overall distortion is given by

$$\bar{x}_k = \bar{x}_r + \bar{x}_g \tag{41}$$

Finally, the image coordinates are obtained from

$$\begin{bmatrix} u_c \\ v_c \\ 1 \end{bmatrix} = \begin{bmatrix} f_{cx} & 0 & u_{0c} \\ 0 & f_{cy} & v_{0c} \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} x_k \\ y_k \\ 1 \end{bmatrix} \tag{42}$$

where $\bar{f}_c = [f_{cx}, f_{cy}]$ is the focal lengths and $\bar{p}_{0c} = [u_{0c}, v_{0c}]$ is the principal point coordinates. Therefore, the model parameters are $\mathcal{L}_c = \{\bar{f}_c, \bar{p}_{0c}, \bar{k}_c\}$.

Depths Camera Calibration (intrinsic) We use same notations that before, except that the index c is replace by d, the depth. Therefore, there are the intrinsic parameters $\bar{f}_d$ for the focal and $\bar{p}_{0d}$ for the principal point of the depth camera. We recall that a depth camera (in particular the kinect) returns a disparity value for each points. A disparity represent the distance that a point moves from the original image (emitter) to the depth camera (receiver). An example to illustrate disparity measurement consists in closing each eye alternatively. In that case, points near of the eyes would move more that further points. Therefore, the depth is given by

$$z_d = \frac{1}{c_1 d_k + c_0} \tag{43}$$

where c_0 and c_1 are intrinsic parameters to be calibrated and d_k is the disparity after distortion correction. Using the above equation for calibration would lead poor result. Indeed a fixed error pattern is observed. In [16], it is proposed to use a spatially varying offset Z_δ such that

$$z_{dd} = z_d + Z_\delta(u, v). \tag{44}$$

Nevertheless, in the present article, a more accurate calibration is designed by correcting the distortion directly on the disparity units. The model is inspired by empirical observations whose proposed an exponential decay (? not totally clear for me ?). For a disparity d given by the kinect, the corrected disparity, d_k, is defined by

$$d_k = d + D_\delta(u, v) \exp(\alpha_0 - \alpha_1 d) \tag{45}$$

where D_δ contains the spatial distortion pattern and $\alpha = [\alpha_0, \alpha_1]$ are intrinsic parameters modeling the decay of the distortion effect with respect to the disparity. The present equations represent the backward model (e.i. measure disparity transform to metric coordinates). The forward problem is obtained by computing the inverse of the equations. We get

$$d_k = \frac{1}{c_1 z_d} - \frac{c_0}{c_1} \tag{46}$$

and

$$d = d_k + \frac{W(\alpha_1 D_\delta(u, v) \exp(\alpha_0 - \alpha_1 d_k))}{\alpha_1} \tag{47}$$

where W is the Lambert W function (the inversion needs some tricks). Finally the model parameters for the intrinsic calibration of the depth camera is given by $\mathcal{L}_d = \{\bar{f}_d, \bar{p}_{0d}, \bar{k}_d, c_0, c_1, D_\delta, \alpha\}$. The three firsts parameters are similar to the one in the color camera calibration.

Extrinsic and Relative Pose We recall that for the calibration, we use a dashboard for the RGB camera (calibration pattern) and a calibration plane for the depth camera. Recall that the checkerboard has to be coplanar with the calibration plane (dashboard) as represented in figure 15. The calibration is done using Zhang's method [4]. Nevertheless, the relative pose of the RGB camera is given with respect to W (the dashboard) and the one for the depth camera is given with respect to V (the plane reference). The transformation between V and W is a priori unknown.

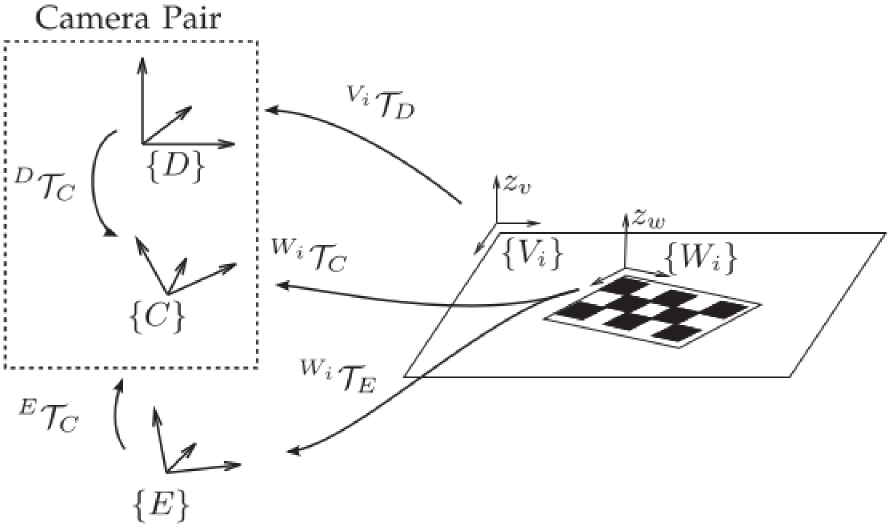


Fig. 4. Reference frames and transformations. $\{D\}$, $\{C\}$, and $\{E\}$ are the depth, color, and external cameras, respectively. For image i, $\{V_i\}$ is attached to the calibration plane and $\{W_i\}$ is the calibration pattern.

Figure 15: Reference frames and transformations. Recall that in this model an external color camera is also considered.

Calibration Method For the RGB camera, we proceed as in the Zhang's method [4] where the corners of the checkerboard (damier) are extracted in image. The RWC of the corners are known. The distortion parameters $\bar{k}_d = [k_1, ..., k_5]$ are initially set to zero. The same method is used for the depth calibration. However, since the checkerboard is not visible, we extract the corner of the dashboard (calibration plane). With this method we can obtained the intrinsic parameter $\bar{f}_d$, $\bar{p}_{0d}$ and the rigid transformation ${}^{V_i}\mathcal{T}_D$; more precisely, we obtain an *initial* guest for the intrinsic parameters. The equation of a plane is given by $\bar{n}^T x - \delta = 0$. In our model, we will use the fact that the checkerboard and the dashboard are coplanar. That is that the equation of the reference plane, in both referential, is given by $\bar{n} = [0, 0, 1]^T$ and $\delta = 0$. Therefore for a rigid transform $\mathcal{T} = \{R, t\}$ (where $R = [r_1|r_2|r_3]$), the normal plane is given by

$$\bar{n} = r_3 \text{ and } \delta = r_3^T t. \tag{48}$$

In order to obtain the relative pose of the depth camera with respect to the color camera, we use several images (n) and defined the concatenated matrices $M_c = [\bar{n}_{c1}, ..., \bar{n}_{cn}]$, $b_c = (\delta_{c1}, ..., \delta_{cn})$ and likewise for the depth camera to form M_d and b_d. Finally, the relative transformation is given by

$$
\begin{aligned}
{}^{C}R'_D =& M_d M_c^T, & (49)\\
{}^{C}t_D =& (M_c M_c^T)^{-1} M_c (b_c - b_d)^T. & (50)
\end{aligned}
$$

Due to the noise, the matrix ${}^{C}R'_D$ is not necessary orthonormal. A valid rotation matrix is obtained through SVD. One have ${}^{C}R_D = UV^T$, where USV^T is the SVD of ${}^{C}R'_D$. The above procedure leads a good calibration for the color camera but a poor one for the depth camera (only gives a good initialization guest). In order to improve the calibration, we have to do some nonlinear minimization (Levenberg-Marquardt). We denote by $\hat{p}_c$, $\hat{d}$ and $\hat{d}_k$ the observed corner positions in the image. The cost function to minimize is given by

$$c = \frac{1}{\sigma_c^2} \sum \|\hat{p}_c - p_c\|^2 + \frac{1}{\sigma_d^2} \sum \|\hat{d} - d\|^2 \tag{51}$$

where the variables σ_c and σ_d correspond to the variance of the measurements. Nevertheless, the above cost function is not very convenient because it involves the evaluation of the Lambert W function at each iterations. Therefore, they propose to use the following cost function instead

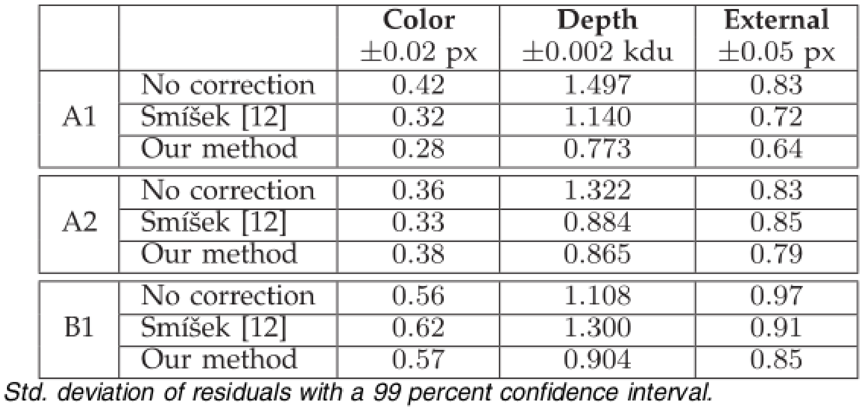
$$c = \frac{1}{\sigma_c^2} \sum \|\hat{p}_c - p_c\|^2 + \frac{1}{\sigma_d^2} \sum \|\hat{d}_k - d_k\|^2. \tag{52}$$

During this step, the spatial disparity pattern D_δ is constant (recall that $\#D_\delta = 640 \times 480 = 307200$). The initial values for α and D_δ are set to zero. Finally, the spatial disparity pattern is obtained, in a second step assuming, by assuming the other parameters to be fixed. Therefore we define the cost function

$$c_d = \sum_{\text{images}} \|\hat{d}_k - d_k\| = \sum_{\text{images}} \sum_{u,v} (\hat{d} + D_\delta(u,v) \exp(\alpha_0 - \alpha_1 \hat{d}) - d_k)^2. \tag{53}$$

It is said that the above equation is quadratic in each $D_\delta(u,v)$ and hence the optimal value of each $D_\delta(u,v)$ is obtained by solving a linear equation (I think that the linear equation is obtained by computing the derivative and equalizing to zero).

Results In figure 16, one can see that the proposed method is the best.

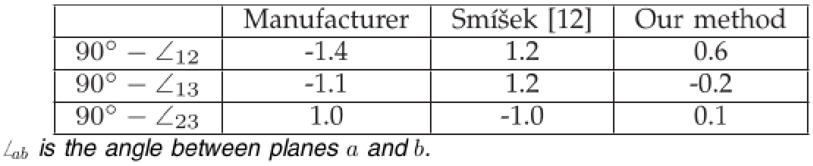


		Color ±0.02 px	Depth ±0.002 kdu	External ±0.05 px
A1	No correction	0.42	1.497	0.83
	Smíšek [12]	0.32	1.140	0.72
	Our method	0.28	0.773	0.64
A2	No correction	0.36	1.322	0.83
	Smíšek [12]	0.33	0.884	0.85
	Our method	0.38	0.865	0.79
B1	No correction	0.56	1.108	0.97
	Smíšek [12]	0.62	1.300	0.91
	Our method	0.57	0.904	0.85

Std. deviation of residuals with a 99 percent confidence interval.

Figure 16: Standard deviation with a 99 percent confidence interval. The calibration is done using the external camera as well. The propose model is the best.

On figure 17, one can see that the calibration method give good result for the angle between normal plane. The reference model is shown in figure 18.



	Manufacturer	Smíšek [12]	Our method
$90° - \angle_{12}$	-1.4	1.2	0.6
$90° - \angle_{13}$	-1.1	1.2	-0.2
$90° - \angle_{23}$	1.0	-1.0	0.1

$\angle_{ab}$ is the angle between planes a and b.

Figure 17: Error between reconstruted planes. The references planes are given in figure 18.

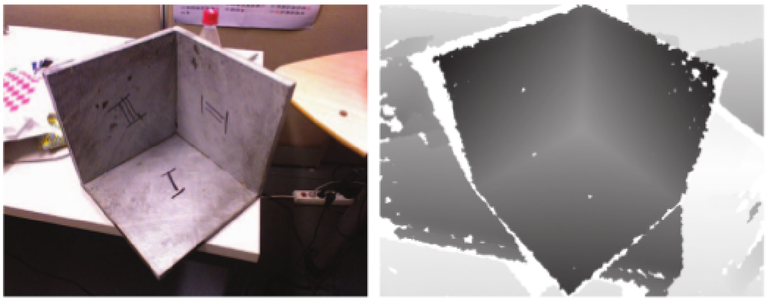


Figure 18: Reference Cube.

7 Gaze experiments

In the present section, I will present a protocol of experiences in order to use (leverage) gaze informations in the learning by demonstration of the robot.

7.1 Materials

We suppose to do the experiment with the BAXTER robot. BAXTER has a suction cup in the left hand and a clamp for grasping in the right hand. An interseting aspect of BAXTER is that it is a human-like robot. The head is the screen of a tablet. Even if eye movement of the robot can be shown on the screen, head movement may be less realistic (interaction). We would fix a kinect camera on the head of BAXTER (directly over the screen or on the component that the screen is attached). An other possibility is to fix the kinnect to the body (chest) of BAXTER, in order to get a frontal view (see figure 22). In that case, we may have a more accurate gaze tracker since eyes are more visible and the frontalization may produce less error. BAXTER has an integrated RGB camera (in the tablet). In order to catch multimodal data, we will also need to use microphone (one or two). Finally, in order to integrate the interaction between human and robot, we will need to use a speaker reproducing the voice of BAXTER.
We want to observe how well we can catch the intention of a teacher (relevant aspect of the task) in various task with various objects. The baseline of the experiment will consist of a table with different object on it and with the robot on one side and the human teacher on the other side. The distance between them will correspond roughly to 1.0-1.5 meter. We list object of interest below. We suppose to have at least two object of each kind (small and big). The different objects that may be relevant to manipulate are

- Cube: The cube is symmetric object, its orientation does not matter. Nevertheless, placing the small cube on the big one is easier that the opposite where we have to care about equilibrium.
- Parallelepiped: We could use parallelepiped with a big difference between width and height ($h >> w$). In that case, the orientation of the object is relevant (vertical vs horizontal).
- Ball: The ball is symmetric as the cube. Nevertheless, it does not have any orientation (the cube as to be on one of its face, the ball does not have). Moreover, the ball is a more dynamic object, it may move after placing it.
- LEGO: Using LEGO may be interesting for the design of a task with many sub-task (construction). Moreover, the experience may be reproduce easier.
- Mug and kettle: This is are daily objects. They can both contain a liquid where we expect the robot to act differently.
- Any objects of common life: We expect the robot to be able to use daily objects. We will need to test our method on a more realistic case.

7.2 Camera Calibration

We will need to proceed to a calibration of the cameras (RGB of kinnect, depth of kinnect and RGB of BAXTER). The calibration method would be the Herrera method [5] and Zhang method [4]. We will need a checkerboard on a dashboard (the checkerboard is not visible from the depth camera). If the intrinsic parameters are already calibrated (should be the case, need to find the information), only the extrinsic parameters (RWC of the cameras, rigid transform) have to be calibrated. Nevertheless, I suppose that the extrinsic calibration parameters of the kinect RGB and depth camera is already calibrated. In that case, we have to calibrate the BAXTER RGB camera with the RGB camera of the kinect. This can be done with the MATLAB toolbox.

7.3 Gaze

Gaze is an important cue in non-verbal communications [17]. It is also very important in social interaction. There are various types of gaze [18]:

- Mutual Gaze: when two people are looking at each other. Usually when they are talking to each other. They might look at eyes or, more generally at face. Indeed, we usually saccade from eyes area to mouth area (and this is person-specific as well, need to find the reference).
- Referential Gaze (or Deictic Gaze): gaze directed to an object or location space. We will certainly analyze this kind of gaze for pick and place example. Gazing at an object may be ok but recognizing a gaze to a location may be more difficult. In the first case, the gaze might coincide with and object position (given by the object recognition module) but in the case of location space we do not have a reference. We need to know the state of the gaze (e.i. saccade vs fixation). In the case of fixation, we should decide if there is an object or if we look at a location space. If location space, what is this location (mean and precision).
- Joint attention: when two agents are looking at the same location or object or person.
- Gaze aversion (saccade): gaze shift away from the main direction of gaze. Usually correlated to cognitive effort [19].

7.3.1 Gaze for conversation and speech

Mutual gaze is an important aspect in conversation and speech. More informations and references are given in [18, Section 3.1]. Some important aspect are listed below.

- A listener would look at the person being listened (the speaker) 88% of the time [20].
- A speaker would look at the target of its speech 77% of the time [20].
- Listener-directed gaze may occure less frequently than speaker-directed gaze. We look more often to a speaker than the speaker look at us. [21].
- More interesting aspect is the *reference action sequence* (when a teacher refers to an object and then the learner acts on that object, +- Rosalis topic). In that case, we can divide the process in five cyclically phases (each with there own distinct gaze behaviors): pre-reference, reference, post-reference, action and post-action [22]. The learner will follow the gaze of the teacher in the firsts phases and then the role are reverse.

- Gaze behavior is person specific. In particular, [23] shows how extroverts and introverts personality change the gaze behavior during conversation. Moreover, this behavior does not only depend independently on each individual trains but depends on the interpersonal dynamics of the partners.

7.3.2 Gaze for object reference and manipulation

The gaze gives a good information about the object of interest and the intention of the user. Yarbus shows 60 years ago that saccadic eye movement reflect cognitive processes [24]. Gaze is an important information about the task to perform. In [25], it is shown that the saccades are made almost exclusively to objects involved in task (before instruction 48% of the gaze to irrelevant object, after it drops to 16%). It is also shown that we look at one object at a time (roughly corresponding to the manipulation duration). But the gaze is not fix and may fix different part of the object [25]. The experiment presented in [25] consists in doing a sandwich or preparing a cup of tea. Land observes three steps in the action each separated approximatively by 0.6 second. Those steps are: body movement, then gaze shift, then manipulation. Therefore, the body movement and the gaze give an information about the next task. That is to say that the gaze fixation may gives next region/object of interest. Nevertheless, the gaze dynamic is different when the next region/object is not initially visible (>50 deg from visual axis). In that case the saccade is trunked. There is a first saccade (based on memory) and then a second small saccade before fixation. In particular, we can say that during manual task with a natural behavior (doing a sandwich, a cup of tea), the fixations mean:

- Object fixation:
  - The object is manipulated.
  - If the object is not manipulated, the object is going to be manipulated.
  - If the object is not manipulated, the object is important for a future task.
- Space location fixation:
  - If the object is manipulated, next reaching location for placing the object.
  - If the object is not manipulated, future relevant space location (an object may be place here not necessary the next object but it may be more expected to be the next one).

An other interesting aspect of gaze fixation is reported in [26]. The experience consists in taking an object and placing it on a space location marked by a belt (virtual experience).

- First, the user can choose the order of picking and placing (size irrelevant).
- Second, s/he picks first the taller object and place it on the same belt (size relevant for pick).
- Third, s/he picks the taller object and place it a given belt. Then, s/he picks the smaller object and place on an other belt (size relevant for pick and place).

The experience consists in changing the size of the object during the manipulation and reporting if the change has been observed. In the case where size was irrelevant, the change of size has not been observed. This is not directly linked to our topic but the experience is interesting for its gaze behavior. An aspect which is not analyzed in [27] is the gaze pattern during a fixation. As mentioned, before [25], we never have a totally fixed gaze during fixation. In particular, we look at different part of the object. It might be interesting to see if this pattern depend on the size (or shape) of the object. Moreover, if it change when size (or shape) is relevant vs irrelevant.

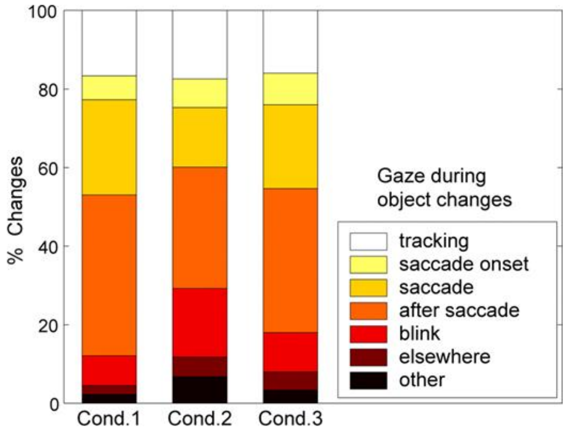


Figure 19: Gaze state during the experience done in [27] and reported in [26].

In figure 19, the gaze behavior is reported. One can observe that no significant changes are observable (maybe more blink in condition 2 and 3 because of the cognitive aspect). Moreover, in [27], the fixation time is reported when no change occurred, unnoticed change and noticed change. We observe that the fixation time significantly increase when the change is noticed. Nevertheless, the gaze behavior is more or less the same in the three conditions.

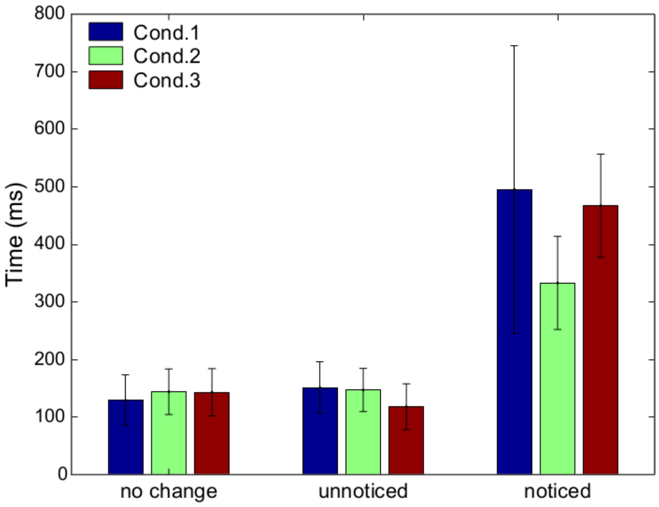


Figure 20: Summed times spent fixating the brick during put-down when either no change occurred, a change was unnoticed, or the change was noticed, for the three different task conditions. Error bars indicate standard error of the mean.

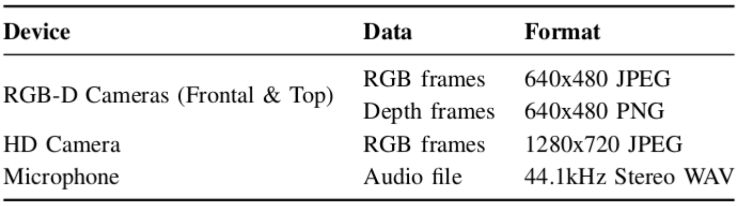
7.4 Unsupervised Gaze Calibration

The model for gaze tracking follows the method presented in [3] and it is based on the works [12, 28, 13].

Dataset Five datasets are available for testing the gaze tracking methods.

- EYEDIAP dataset [29] consists of sitting people looking at a target (ball with suspending by a rope) moving in front of them. Different set-up are available fixed head vs moving head. The dataset does not take place in a natural behavior
- SONVB [30]: job interviews with annotation.
- UBIimpressed dataset [31]: interview situations and desk situation (free to move).
- VERNISSAGE dataset [32]: No depth camera but vicon sensors. Can not be used with the head traking based on 3D morphable model [12, 3]. In [1], they use head tracker defined in [8] (based on texture and color mapping).
- MHRI dataset [33]: The experiment is done using BAXTER. It has multimodal data (rgb, depth, audio). The experiment has three interaction type
  - Point: "This is a box" while pointing to the box.
  - Show: "This is a box" while holding the box.
  - Speak: "The box is next to, has something on top, it is on top of, ..."

  There are 22 objects corresponding to to common life object (mug, banana, knife, bottle,...). The experiment is conducted with 10 participants with 30 interaction each (10 of each type). The position of the cameras and microphone is shown in figure 22. We will certainly use an equivalent setup. Nevertheless, we will certainly use one kinect. I should check which one gives good results. The format specification for the kinect and the rgb camera are given in figure 21. The contact mail is `pazagra@unizar.es`.



Device	Data	Format
RGB-D Cameras (Frontal & Top)	RGB frames	640x480 JPEG
	Depth frames	640x480 PNG
HD Camera	RGB frames	1280x720 JPEG
Microphone	Audio file	44.1kHz Stereo WAV

Figure 21: dataset format specification.

7.4.1 Gaze Calibration Method

In the present section, we will present a general method for calibration of the gaze for user-specific. Let g_t denotes the gaze direction at time (frame) t (i.e. $g_t = (\phi_t, \theta_t)$, tilt and pan angle). One can define a gaze direction unitary vector through a transform denoted by $\Phi(g_t) = v_t$. In order to make the calibration of the gaze, we define the gaze calibration function $G(v_t, w)$ represented the calibrated gaze direction unitary vector where w represents the calibration parameters. The target unitary vector is denoted by u_t. Similarly, one can define a target direction $\tilde{u}_t = (\tilde{\phi}_t, \tilde{\theta}_t)$ denoting the tilt and pan angle of the target and a calibration gaze direction map $\tilde{G}$ where $\tilde{G}(g_t, \tilde{w})$ represent the calibrated gaze direction and $\tilde{w}$ the calibration parameters. With those notations, one can define the calibration process as an optimization problem. Indeed, the goal is to minimize the

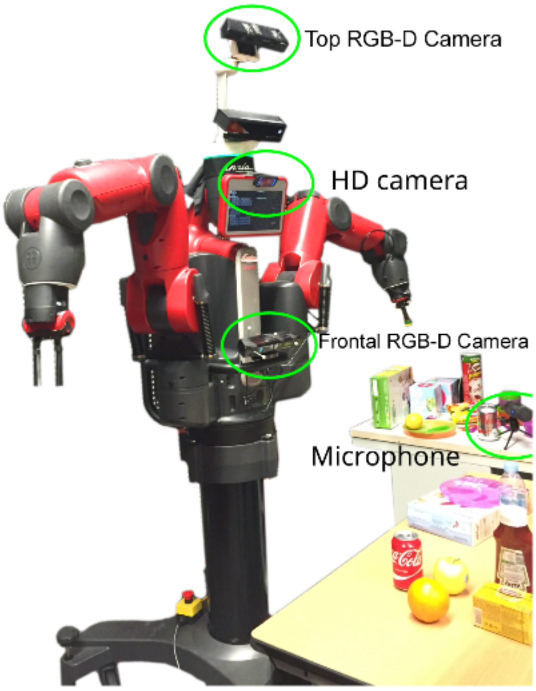


Figure 22: BAXTER robot used to acquire the MHRI dataset.

difference between $\tilde{u}_t$ and $\tilde{G}(g_t)$ for a set of samples (batch minimization). In L^2-norm, this gives

$$\min_{\tilde{w}} \sum_{t=1}^{n} \|\tilde{u}_t - \tilde{G}(g_t, \tilde{w})\|_2^2 \tag{54}$$

Choosing the following calibration function

$$\tilde{G}(g_t, \tilde{w}) = g_t + \tilde{w}, \tag{55}$$

the solution of the L^2-norm minimization problem is given by

$$\tilde{w} = \frac{1}{n} \sum_{t=1}^{n} (g_t - \tilde{u}_t). \tag{56}$$

It corresponds to the mean average calibration presented in [3]. Considering the L^1-norm minimization

$$\min_{\tilde{w}} \sum_{t=1}^{n} \|\tilde{u}_t - \tilde{G}(g_t, \tilde{w})\|_1, \tag{57}$$

with the same calibration function (55). The solution is given by

$$\tilde{w} = \text{median } \{g_t - \tilde{u}_t \ : \ 1 \leq t \leq n\}. \tag{58}$$

One get the median average calibration. In order to get the Least median of squares estimator [34], one have to solve the minimization problem

$$\min_{\tilde{w}} \operatorname*{med}_{t=1,\dots,n} \ \|\tilde{u}_t - \tilde{G}(g_t, \tilde{w})\|^2 \tag{59}$$

with the same calibration function (55). Then, we get the LMedS estimator used in [3]. More precisely, the LMedS method used in [3] use 50% of the median in order to find a least square approximation of the gaze bias. In other words, the LMedS is used to avoid outliers (could we use a threshold?).

Remark The calibration method presented above may be improved in to aspect. First, using a more sophisticated calibration function (polynomial, scaling, ...). In [3], they have observe constant bias, this is why they use the linear calibration function (55). Nevertheless, in there experiment, the subject was looking at a unique (almost fixed) target, the interviewer (or resp. the participant). One should verify if a constant gaze bias is a good model for a more dynamic set-up. For example, the EYEDIAP datset with moving target (fix and moving head motion). An other aspect is the shape of the object we are looking at. In [3], the shape is constant. It is a face. The subject may look between the eyes, the nose, the mouth but it will always be the "same" gaze behavior. This lead us to the second way to improve the method for unsupervised gaze calibration. We need to define a good estimation of u_t (or $\tilde{u}_t$), the reference gaze direction (true gaze). That is to say, we need to define what is u_t in an unspuervised fashion. In [3], they use the fact that we look at the speaker, in particular at the beginning of his speech, afterwards, there is usually a gaze aversion [18, 20, 22]. In our case, the speaker is the robot. If we have a human-like robot, we could use the same principle. That is to say that the target is given by the position of the robot face. The robot has to speak with a human and used the first data for its calibration (+- 1s after the start of the speaking [25] and during +- 1,2s before a gaze aversion, need to find a reference, I read it somewhere). On the other hand, the speaker is looking at the listener (+- 2s after the start of the speaking (because cohnitive gaze aversion) during +- 2 second before a gaze aversion [35]). In [3], they use "only" 20 frames for the gaze calibration. We could expect from a good gaze behavior model that we can extract reference gaze (u_t) with a good confidence.

7.5 Experiments

The experiments will be done with the BAXTER robot. The robot is standing in one side of a table while the teacher is on the other side. They are facing each other (similar to the MHRI dataset). The experiments will involve various object and task. The goal is to extract relevant information about the object (affordance) and the task (goal) from the gaze behavior (but also on other informations). The accuracy of the actual model for gaze tracking is around 6 degrees (interview) and around 10 degrees (desk). This corresponds to an error of 10% (resp. 17.5%) of the distance in each direction (for 1m, we have 20cm (resp. 35 cm) of error). Therefore, we cannot only use the gaze and we should use head pose as well. Due the accuracy of the gaze, it is difficult to extract a precise region where the teacher is looking at. Nevertheless, we could extract the state of the gaze (aversion, fixation, pursuit, saccade, mutual gaze). Knowing those states will inform the robot on the different aspect of the task.

7.5.1 Objects tracking

The robot has to be able to track the different object involve in the experiment. We can use depth camera in order to remove the background and the table. Depending on the view point, we should have a good segmentation of the objects. In the method used in [33] in pair with object recognition method (color histogram, SVM, Nearest neighbors). They use RGB-D data from the Washington dataset (REFERENCE). There is also the YOLO method (REFERENCE) using RGB images and tracked object in real-time. One could use object detection in RGB camera and then use the extrinsic parameter in order to locate the object in the depth camera and provide the position in a 3 dimensional coordinate system.

7.6 TODO list

- Extract gaze from MHRI dataset and use it instead of the hand and pointing tracker. Need gaze calibration.
- Extract head pose and arms skeleton from MHRI dataset.
- Object detection from MHRI dataset. Could use YOLO and depth extrinsic parameters.
- Test gaze accuracy for EYEDIAP dataset. Following the ball and gaze calibration.
- Define an experiment.
- Define gazing point. It is the nearest point to the gaze vectors of each eyes (or a more sophisticate one). May work as a filter.
- Define a multimodal model with gaze, head pose, skeleton and context for a task representation. Extract referent object and placing location.

8 Gaze Calibration Experiments

The goal of this experiment is to define a better person specific gaze calibration. The final objective is to proceed to this calibration in an unsupervised manner. In order to better understand how the gaze can be well calibrated, we will try to answer four questions:

- Should we define some points of fixation or a pursuit?
- Does the shape or size of the object matter for fixation (point vs surface)?
- Does the calibration change between vertical and horizontal region?
- How the calibration will depend on the position of the participant?

8.1 Q1: Fixation vs Pursuit

It has been observed from the existing data that the gaze during the fixation change compare to the pursuit. In particular, the data looks more smooth during a pursuit. Therefore, it would be interesting to know which one between fixation or pursuit would give the better calibration data.

8.2 Q2: Shape and Size

When someone look at a small point, we normally have small variation in the gaze. Nevertheless, when we are looking at a more complex object, for example a face, we will not fix a single point. Therefore, it seems that the fixation point may depend on the shape and the size of the object. Moreover, it may depend on the dynamics of the object, for example, we may look differently to someone who is speaking than someone who do not.

8.3 Q3: Vertical vs Horizontal

We are interesting to know how this calibration of the gaze depends on the direction we are looking at. That is to say, is a calibration done in the vertical region (robot) suitable for experiment where the gaze is oriented threw an horizontal plane (table) and vice versa?

8.4 Q4: Moving Participant

We would like to know how the gaze calibration may change with respect to the position of the participant. The gaze tracking is based on a frontalization of the head. This frontalization will depend on the head pose with respect to the robot. Therefore, we would like to know how the gaze calibration may be affected by the position of the participant with respect to the robot.

8.5 Remark

An other aspect I wanted to test is the gaze calibration when the head is fixed and when the person is free to move his head. Finally, I have decided to avoid this experiment for two reasons. First, because that would involved to do more experiment and this might be tiring for the participants and for us (data annotations). Secondly, because forcing the participant to not move there head would be to artificial. Therefore, calibration for this kind of artificial set-up may be useless in real case.

8.6 Set-up

The set-up is defined in the section concerning the materials. For this experiment, we will use the kinect with frontal view and the RGB camera in the head of BAXTER. Therefore, we will need to proceed to a calibration of the camera using Zhang method [4]. Moreover, we will need to proceed to the calibration of the robot with respect to the camera. This can be obtained by moving the end-effector in front of the camera and use ROS to extract the end-effector position (use of a marker). Manu has the code and method for that. We will need 10 participants for the experience. Actually, we do not need a lot of participant since we want to test a calibration method for each participant. We are not learning a general method. In other words, we want to over-fit the data for each participant. It would be interesting to have different age, sex and ethnicity. Moreover, it would be interesting to have participant with glasses.

8.7 E1: Using end-effector of BAXTER

In this experiment, we will use the BAXTER robot and we will use the fact that we know the position of the end-effector. Therefore, we will not need to do the data annotations. In order to make BAXTER look like a human, we will display a face on the tablet. The experiment will begin with BAXTER saying "Hello! how are you today?" while he is raising its right hand (like he says hi). Then, he says "Please, fix my right hand and follow it". Then, he will do the same with the left hand. The motion of the end-effector is random. It will consist of some point in space that he has to reach and then stay in that position for 3 seconds. For each arm, we choose randomly 5 points in the reachable region of the robot. Once, the robot has reach the five points, it comes back to the initial position and make an horizontal "8" with its arm.

8.7.1 Remark

Here is a list of remark for this experiment:

- We choose the points randomly in order to make variation between the samples. We do not want over-fitting. We do want over-fitting for each participant but we do not want over-fitting between the participant.

- We do the experiment with the two arms in order to catch the difference in the calibration between the left and right region. Moreover, due to the position of the camera, we may have some occlusion and therefore calibrations would not be possible.
- I choose 3 seconds for the fixation because, we should avoid the 0.6 first and last second of the fixation [25]. Therefore, we should have more or less 2 second of fixation. In [3], they use 20 frames for calibration. At 30 fps (kinect), we should have enough data even during 1 second.
- I choose 5 points for each arm which will correspond to 10 reachable regions for the robot. I suppose that the data may last more or less 1 minute for each arm. (10 second for interaction part, 30 second for calibration with fixation, 20 second for dynamic calibration).
- The choice of the reachable regions may be chosen by considering the occlusion of the camera or not (I have to take a decision).
- In this experiment, the fixation point is not explicitly given. We suppose that people will fixed the end-effect but we are not sure where exactly and if it will be always the same point. Nevertheless, I suppose that the variation would be small since the end-effect is a small region. Moreover, it might be interesting to consider a natural fixation of the end-effector since in E2 we are considering more artificial fixation points. Nevertheless, we can also make BAXTER hold an item with a clear target on (ping-pong ball with a dot).

8.7.2 Calibration and Annotation

The intrinsic and extrinsic parameters of the kinect should be available in the Idiap database. If not, we will use Herrera method. We will need to proceed to a time calibration between the robot and the camera. One possibility is to timestamp each frame with ROS (as done in [33]).

8.7.3 Test set

In order to test our model, I was thinking to use and n-fold cross-validation method. We will also test the calibration done with the fixation on the pursuit data and vice versa (Q1). We will also test the calibration with the table fixations (E2 and E3) in order to answer to Q3 and Q2.

8.7.4 Relevance

In this experiment, we aim to test the calibration method for fixation and pursuit. This experiment may be useful to answer the question Q1. Moreover, it will give a database for the calibration in the vertical region (robot). It will therefore be used to answer question Q3 as well.

8.8 E2: Table fixation

We will put 5 markers (small points) on the table (one in each corner and one in the middle). We will ask the participant to look at each point in a given ordering (we will number the point). We will tell the participant when he can move to the next point.

8.8.1 Remark

Here is a list of remark of this experiment:

- In this experiment only fixation is considered.
- We will change the ordering of the target point between participant.
- Maybe we should consider more than 5 fixation points. (I am not sure)

8.8.2 Calibration and Annotation

The position of the target points can be obtained with image processing. We can extract the five points from the image by annotation by hands (the points are fixed in each frame since BAXTER is not moving) or using automatic detection. In order to obtain a RWC of the points, we will measure the distance between them (coordinate in the table reference). Then, using calibration method, we will be able to predict the coordinate with respect to the kinect. In this experiment, we do not need to calibrate the robot but we have to know when the participant is fixing an object. Therefore, we have to define an efficient annotation. For each participant, we know the ordering of the fixation (we will say which object to look and when to do it). Nevertheless, we do not have access to the exact time he is looking at the object. We could imagine two annotation methods:

- Use a method that extract fixation vs pursuit (Remy is working on that). Since we know the ordering of the fixation, we will be able to annotate the data.
- Use a calibration file where the timestamped of the fixation is saved in live (we push a button when we ask the participant to look at an other target). We will need to consider only the middle part of the fixation as in E1.

8.8.3 Test Set

In order to test the model, we will use a 5-fold cross-validation method. We can also test our calibration method using the data of E1 and give an answer for Q3 and Q1. We can also test our calibration on the data of E3 and answer to Q2.

8.9 E3: Table object fixation

We will place five objects on the table. The participant has to look at them in a given order. We will say when he can move to the next object. The object used are presented in the materials section. We will change randomly the objects involved in the experiment.

8.9.1 Calibration and Annotation

For the calibration in time, we will proceed as in E1. Concerning the annotation, we will need to define a point of fixation for each object. Since the object are placed on one of the five marked placed. We can define the position of the object. We will need to annotate at which place is each object.

8.9.2 Test Set

We should proceed as in E2 for the validation of the method.

8.10 E4: Moving Participant

We will place some points on the floor were the participant has to stand. Then BAXTER will ask a question to the participant and he has to answer. Then, the participant has to move to an other location and BAXTER will ask an other question. We could define 5 locations (three directly in front of the table and two bit further). In order to make the interaction natural, we will need to display a human kind face on the tablet. Moreover, it would be useful to display a natural gaze behavior and mouth behavior (I had to look for existing softwares).

8.10.1 Calibration and Annotation

We will use the extrinsic calibration of the kinect and the RGB camera in order to localize the tablet and, therefore, define the ground truth. We will need to calibrate in time the question asked by the robot and the kinect input. Since the interaction is natural, we could test Remy's method to identify fixation. We can also proceed to annotations by hand.

8.10.2 Test Set

In order to validate the calibration, we can proceed by cross-validation. We can also use the other experiments for testing the calibration.

8.11 E5: Natural Behavior

In this experiment the participant has to take an object and place it somewhere else. Different set-up may be interesting (placing an object on an other, moving an object with an other on it,...). I will work on this later. I believe that this experiment may be useful for gaze calibration in natural behavior and also for leveraging the gaze behavior in a task understanding.

8.12 TODO list

- Check-out ROS: how data are obtained, access the different topics, calibrate timestamps.
    - Access kinect.
    - Access RGB camera.
    - Access end-effector.
    - Access speaker.
    - Access tablet screen.
- Get a kinect and fixed rigidly to BAXTER.
- Get the speaker.
- Look for virtual human agent (display on the tablet).
- Prepare the codes for calibration of the robot and the cameras (see with Manu).
- Control the robot end-effector.
- Prepare the different objects (cube, parallelepiped,...).

- Prepare the code for gaze tracking. In particular, give the correct inputs (calibration) and check how well it works for real-time computation.

9 Robot and Camera Calibration

Let us denote the camera frame by $\mathcal{F}_X$ and the robot frame by $\mathcal{F}_Y$. Let us denote by X and Y the position of the end-effector in, respectively, the camera and robot frame. Moreover, we define X^m and Y^m as the position of the marker in,respectively, the camera and robot frame. We seek for a rigid transform $(\bar{R}, \bar{T})$ such that

$$\bar{R}Y + \bar{T} = X \text{ and } \bar{R}Y^m + \bar{T} = X^m \tag{60}$$

The pose of the end-effector can be characterized by its orientation, Q_Y, and position Y. Therefore (Q_Y, Y) defines a frame. The marker can be represented in this frame by a simple translation vector denoted by t_Y (the marker has the same orientation than the frame). That is that the position of the marker in the end-effector frame is t Hence, the marker frame is given by $(Q_Y, Y + Q_Y t)$. There is also a rigid transformation between the end-effector in the robot frame and the marker in the camera frame. Let us denote this transformation by

$$RY + T = X^m. \tag{61}$$

The solution of the above transformation is very similar to the procustes problem can be computed easily. In order to obtain the rigid transform $(\bar{R}, \bar{T})$, we proceed as follow

$$\bar{R}Y^m + \bar{T} = X^m \Rightarrow \bar{R}(Y + Q_Y t) + \bar{T} = X^m \Rightarrow \bar{R}Y + \bar{R}Q_Y t + \bar{T} = X^m \tag{62}$$

For a set of sample $Y = (Y_1, ..., Y_n)$ and $X^m = (X_1^m, ..., X_n^m)$, The equation reads

$$\bar{R}Y_i + \bar{R}Q_{Y_i} t + \bar{T} = X_i^m \tag{63}$$

By subtracting the mean for both side. we get

$$\bar{R}(Y_i - \mathbb{E}[Y]) + \bar{R}(Q_{Y_i} - \mathbb{E}[Q_Y])t = X_i^m \mathbb{E}[X_i^m] \tag{64}$$

Hence

$$\bar{R}\bar{Y}_i + \bar{R}\bar{Q}_{Y_i} t = \bar{X}_i^m \tag{65}$$

9.1 Algorithm

1. Collect synchronized data: X_i^m and (Q_{Y_i}, Y_i) for $i = 1, ..., N$.

2. Compute

$$\bar{X}_i^m = X_i^m - \mu_{X^m} \text{ with } \mu_{X^m} = \frac{1}{N}\sum_{i=1}^{N} X_i^m \tag{66}$$

and

$$\bar{Y}_i = Y_i - \mu_Y \text{ with } \mu_Y = \frac{1}{N}\sum_{i=1}^{N} Y_i \tag{67}$$

3. Gradient descent. Get $\bar{R}$ and t. Set $\bar{T}$

10 Meeting, 14 August

- Inverse kinematics: modify gradient descent (damping),
- Robot calibration: Use orientation to get relative position of the marker,
- Time synchronization
- MHRI dataset: calib +- ok, problem with CPI.
- Codes: kinect processing (calib, lossless), Baxter kinematics (IK, FK), Robot-camera calibration.
- EPFL Lecture: Machine Learning (Jaggi Martin, Urbanke Rüdiger)
  1. Basic regression and classification concepts and methods: Linear models, overfitting, linear regression, Ridge regression, logistic regression, and k-NN.
  2. Fundamental concepts: cost-functions and optimization, cross-validation and bias-variance trade-off, curse of dimensionality.
  3. Unsupervised learning: k-Means Clustering, Gaussian mixture models and the EM algorithm.
  4. Dimensionality reduction: PCA and matrix factorization, word embeddings
  5. Advanced methods: generalized linear models, SVMs and Kernel methods, Neural networks and deep learning
- TODO: code for calibration acquisition (press space to save), camera calibration file (vertical, horizontal), first tests.

References

[1] S. Sheikhi and J.-M. Odobez, “Combining dynamic head pose-gaze mapping with the robot conversational state for attention recognition in human-robot interactions.,” *Pattern Recognition Letters*, vol. 66, pp. 81–90, 2015.

[2] B. Massé, S. O. Ba, and R. Horaud, “Tracking Gaze and Visual Focus of Attention of People Involved in Social Interaction.,” *CoRR*, vol. abs/1703.04727, 2017.

[3] R. Siegfried, Y. Yu, and J.-M. Odobez, “Towards the use of social interaction conventions as prior for gaze model adaptation.,” in *ICMI* (E. Lank, A. Vinciarelli, E. E. Hoggan, S. Subramanian, and S. A. Brewster, eds.), pp. 154–162, ACM, 2017.

[4] Z. Zhang, “A Flexible New Technique for Camera Calibration.,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 22, no. 11, pp. 1330–1334, 2000.

[5] D. Herrera C., J. Kannala, and J. Heikkilä, “Joint Depth and Color Camera Calibration with Distortion Correction.,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 10, pp. 2058–2064, 2012.

[6] S. R. Langton, R. J. Watt, and V. Bruce, “Do the eyes have it? Cues to the direction of social attention,” *Trends in Cognitive Sciences*, vol. 4, no. 2, pp. 50–59, 2000.

[7] D. A. Hanes and G. McCollum, “Variables Contributing to the Coordination of Rapid Eye/Head Gaze Shifts.,” *Biological Cybernetics*, vol. 94, no. 4, pp. 300–324, 2006.

[8] V. Khalidov and J.-M. Odobez, “Real-time Multiple Head Tracking Using Texture and Colour Cues,” 2017.

[9] E. G. Freedman and D. L. Sparks, “Eye-Head Coordination During Head-Unrestrained Gaze Shifts in Rhesus Monkeys.,” *Journal of Neurophysiology*, vol. 77, p. 2328, 1997.

[10] E. G. Freedman, “Interactions between eye and head control signals can account for movement kinematics.,” *Biological Cybernetics*, vol. 84, no. 6, pp. 453–462, 2001.

[11] S. O. Ba and J.-M. Odobez, “Recognizing Visual Focus of Attention From Head Pose in Natural Meetings.,” *IEEE Trans. Systems, Man, and Cybernetics, Part B*, vol. 39, no. 1, pp. 16–33, 2009.

[12] K. A. F. Mora and J.-M. Odobez, “Gaze Estimation in the 3D Space Using RGB-D Sensors - Towards Head-Pose and User Invariance.,” *International Journal of Computer Vision*, vol. 118, no. 2, pp. 194–216, 2016.

[13] Y. Yu, K. A. F. Mora, and J.-M. Odobez, “Robust and Accurate 3D Head Pose Estimation through 3DMM and Online Head Model Reconstruction.,” in *FG*, pp. 711–718, IEEE Computer Society, 2017.

[14] R. Hartley and A. Zisserman, *Multiple View Geometry in Computer Vision.* New York, NY, USA: Cambridge University Press, 2 ed., 2003.

[15] J. Heikkilä and O. Silvén, “A Four-step Camera Calibration Procedure with Implicit Image Correction.,” in *CVPR*, pp. 1106–1112, IEEE Computer Society, 1997.

[16] J. Smisek, M. Jancosek, and T. Pajdla, "3D with Kinect.," in *ICCV Workshops*, pp. 1154–1160, IEEE, 2011.

[17] M. Argyle, " Non-Verbal Communication in Human Social Interaction. .," *R. A. Hinde, Non-verbal communication. Oxford, England: Cambridge U. Press.*, 1972.

[18] H. Admoni and B. Scassellati, "Social Eye Gaze in Human-Robot Interaction: A Review," *Journal of Human-Robot Interaction*, vol. 6, no. 1, pp. 25–53, 2017.

[19] S. Andrist, B. Mutlu, and M. Gleicher, "Conversational Gaze Aversion for Virtual Agents.," in *IVA* (R. Aylett, B. Krenn, C. Pelachaud, and H. Shimodaira, eds.), vol. 8108 of *Lecture Notes in Computer Science*, pp. 249–262, Springer, 2013.

[20] R. Vertegaal and Y. Ding, "Explaining effects of eye gaze on mediated group conversations: : amount or synchronization?," in *CSCW* (E. F. Churchill, J. F. McCarthy, C. Neuwirth, and T. Rodden, eds.), pp. 41–48, ACM, 2002.

[21] M. Cook, "Gaze and Mutual Gaze in Social Encounters: How long—and when—we look others "in the eye" is one of the main signals in nonverbal communication," *American Scientist*, vol. 65, no. 3, pp. 328–333, 1977.

[22] S. Andrist, W. Collier, M. Gleicher, B. Mutlu, and D. Shaffer, "Look together: analyzing gaze coordination with epistemic network analysis," *Frontiers in Psychology*, vol. 6, p. 1016, 2015.

[23] S. Andrist, B. Mutlu, and A. Tapus, "Look Like Me: Matching Robot Personality via Gaze to Increase Motivation," in *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, CHI '15, (New York, NY, USA), pp. 3603–3612, ACM, 2015.

[24] Yarbus, "Eye Movements and Vision," *Plenum*, 1967.

[25] M. F. Land, "Vision, eye movements, and natural behavior," *Visual Neuroscience*, vol. 26, no. 1, pp. 51–62, 2009.

[26] M. Hayhoe and D. Ballard, "Eye movements in natural behavior," *Trends in Cognitive Sciences*, vol. 9, no. 4, pp. 188–194, 2005. doi: 10.1016/j.tics.2005.02.009.

[27] J. Triesch, D. H. Ballard, M. M. Hayhoe, and B. T. Sullivan, "What you see is what you need," *Journal of Vision*, vol. 3, no. 1, p. 9, 2003.

[28] K. A. F. Mora and J.-M. Odobez, "Geometric Generative Gaze Estimation (G3E) for Remote RGB-D Cameras.," in *CVPR*, pp. 1773–1780, IEEE Computer Society, 2014.

[29] K. A. F. Mora, F. Monay, and J.-M. Odobez, "EYEDIAP: a database for the development and evaluation of gaze estimation algorithms from RGB and RGB-D cameras.," in *ETRA* (P. Qvarfordt and D. W. Hansen, eds.), pp. 255–258, ACM, 2014.

[30] L. S. Nguyen, D. Frauendorfer, M. S. Mast, and D. Gatica-Perez, "Hire me: Computational Inference of Hirability in Employment Interviews Based on Nonverbal Behavior," *IEEE Transactions on Multimedia*, vol. 16, pp. 1018–1031, June 2014.

[31] S. Muralidhar, L. S. Nguyen, D. Frauendorfer, J.-M. Odobez, M. Schmid Mast, and D. Gatica-Perez, "Training on the Job: Behavioral Analysis of Job Interviews in Hospitality," in *Proceedings of the 18th ACM International Conference on Multimodal Interaction*, ICMI 2016, (New York, NY, USA), pp. 84–91, ACM, 2016.

[32] D. B. Jayagopi, S. Sheiki, D. Klotz, J. Wienke, J. M. Odobez, S. Wrede, V. Khalidov, L. Nyugen, B. Wrede, and D. Gatica-Perez, “The vernissage corpus: A conversational Human-Robot-Interaction dataset,” in *2013 8th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, pp. 149–150, March 2013.

[33] P. Azagra, F. Golemo, Y. Mollard, M. Lopes, J. Civera, and A. C. Murillo, “A multimodal dataset for object model learning from natural human-robot interaction.,” in *IROS*, pp. 6134–6141, IEEE, 2017.

[34] P. J. Rousseeuw, “Least Median of Squares Regression,” *Journal of the American Statistical Association*, vol. 79, no. 388, pp. 871–880, 1984.

[35] S. Andrist, X. Z. Tan, M. Gleicher, and B. Mutlu, “Conversational gaze aversion for humanlike robots.,” in *HRI* (G. Sagerer, M. Imai, T. Belpaeme, and A. L. Thomaz, eds.), pp. 25–32, ACM, 2014.