Are You Talking to Me?
Detecting Attention in First-Person Interactions

Luis C. González-García
and L. Abril Torres-Méndez

Robotics and Advanced Manufacturing Group
CINVESTAV Campus Saltillo
Ramos Arizpe, México
Email: carlos.gonzalez@cinvestav.edu.mx
abril.torres@cinvestav.edu.mx

Julieta Martinez, Junaed Sattar
and James J. Little

Department of Computer Science
The University of British Columbia
Vancouver, Canada
Email: {julm, junaed, little}@cs.ubc.ca

***Abstract*—This paper presents an approach for a mobile robot to detect the level of attention of a human in first-person interactions. Determining the degree of attention is an essential task in day-to-day interactions. In particular, we are interested in natural Human-Robot Interactions (HRI's) during which a robot needs to estimate the focus and the degree of the user's attention to determine the most appropriate moment to initiate, continue and terminate an interaction. Our approach is novel in that it uses a linear regression technique to classify raw depth-image data according to three levels of user attention on the robot (null, partial and total). This is achieved by measuring the linear independence of the input range data with respect to a dataset of user poses. We overcome the problem of time overhead that a large database can add to real-time Linear Regression Classification (LRC) methods by including only the feature vectors with the most relevant information. We demonstrate the approach by presenting experimental data from human-interaction studies with a PR2 robot. Results demonstrate our attention classifier to be accurate and robust in detecting the attention levels of human participants.**

***Keywords*–Human-robot interaction; Body pose classification; Least squares approximations; Raw range data analysis.**

I. Introduction

Determining the attention of people is an essential component of day-to-day interactions. We are constantly monitoring other people's gaze, head and body poses while engaged in a conversation [1][2][3]. We also perform attention estimation in order to perform natural interactions [4][5]. In short, attention estimation is a fundamental component of effective social interaction; therefore, for robots to be efficient social agents it is necessary to provide them with reliable mechanisms to estimate human attention.

We believe that human attention estimation, particularly in the context of interactions, is highly subjective. However, attempts to model it have been relatively successful, *e.g.*, allowing a robot to ask for directions when it finds a human, as in the work of Weiss *et al.* [6]. Nonetheless, the state-of-the-art is still far from reaching a point where a robot can successfully interact with humans without relying on mechanisms not common to natural language. Recently, the use of range images to make more natural human-machine interfaces has been in the agenda of researchers, like in the case of the Microsoft Kinect$^{\text{TM}}$, which delivers a skeleton of a human that can be further used as a high-level feature of the human pose [7]. Although good results have been obtained with such devices in pose estimation, little effort has been devoted to further infer information about the user from such data. In this work, we use range data (similar to that shown in Figure 1) to infer the level of attention of the user, which is not explicitly given by the sensor output.

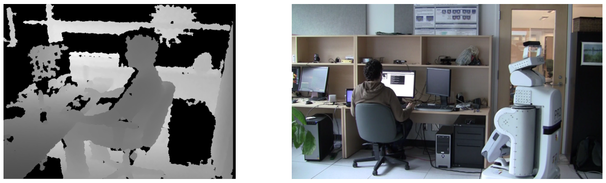


Figure 1. *Left*: Raw range input that a robot gets when trying to asses human attention, as described in this work. *Right*: Set-up scenario for our experiments. The PR2 robot approaches a human sitting at a desk..

Our approach is novel in that it uses raw depth images to evaluate the attention level of a subject, regardless of whether she is facing the depth sensor, in order to classify her pose in an attention scale. In this work, we focus on learning human attention from raw depth images by using the LRC algorithm, which can be exploited by social robots to determine the best moment to ask for support from a human sitting at her desk, like those found in common working or reading spaces.

The remainder of this paper is structured as follows: In Section II, we talk about how other authors have tackled the problem of attention awareness detection using images and range information as a source. In Section III, we describe the problem that this paper faces, attention estimation from a first person perspective using only raw range information. In Section IV, we walk through the technical aspects of the methodology that we propose (LRC). In section V, it is described the actual set-up and execution of the experiments, as well as the interpretation and discussion of the data gathered from them. Finally, in Section VI, our conclusions and suggestions for future work are exposed.

II. Related Work

The problem of attention awareness detection, despite its relevance, remains largely unexplored in the HRI literature. Here we present some of the building blocks of our work.

A. *Pose, Head and Gaze Estimation*

Some of the most effective social cues for attention estimation are gaze, body and head poses. Fortunately, a large body of knowledge has been gathered in these areas.

Shotton *et al.* [7], used a single Red-Green-Blue+Depth (RGB+D) camera to perform pose estimation and body parts recognition. A randomized decision forest was trained on synthetic data that covered a wide range of human poses and shapes. Features were obtained by computing the difference of depth between two points. A further speedup was achieved by providing a GPU implementation. Vision-only approaches range from the Flexible Mixtures-of-Parts [8], an extension of the Deformable Parts Model [9] which explicitly accounts for different body deformations and appearances, to leverage poselet-based part detections for further constraining optical-flow-based-tracking of body parts [10].

Head pose estimation can be seen as a sub-field of full-body pose estimation. In fact, Kondori, Yousefi, Haibo and Sonning [11] extended the work of Shotton *et al.* to Head Pose Estimation in a relatively straightforward manner. Similarly, the problem becomes much harder when depth information is no longer available.

For a more in-depth treatment of the subject, as well as for recent advances in gaze estimation, we direct the reader to the reviews made by Murphy-Chutorian and Trivedi [12] and Hansen and Qiang [13].

B. *Awareness Detection in Computer Vision*

Estimating attention from visual input has been studied particularly in the context of driving. Doshi and Trivedi [14] built a system that incorporated cameras observing both the human subject and her field of view. By estimating the gaze of the subject and the saliency map from her viewpoint, they used Bayes' rule to obtain a posterior distribution of the location of the subject's attention. Our work is different from theirs since just as in person-to-person interactions, we do not have access to the field of view of the person, but we might rather *be* a part of it.

Also related are Mutual Awareness Events (MAWEs). MAWEs are events that concentrate the attention of a large number of people at the same time. In this context, Benfold and Reid [15] built upon evidence from the estimated head poses of large crowds to guide a visual surveillance system towards interesting points.

C. *First-Person Interaction*

Recently, some work has been devoted to transfer knowledge gained from third to first person perspectives. Ryoo and Matthies [16] performed activity recognition from a first-person viewpoint from continuous video inputs. They combined dense optical flow as a global descriptor and cuboids [17] as local interest point detectors, then built a visual dictionary to train an SVM classifier using multi-channel kernels.

D. *Human Attention and Awareness Estimation*

To this day, human attention remains an active area of research. A widely accepted model of attention was proposed by Itti and Koch [18], where attention is understood as the mixture of "bottom-up", *i.e.*, unconscious, low-level features of an image, and "top-down", *i.e.*, task-oriented mechanisms that the subject controls consciously. Later work by Itti and Baldi incorporated the element of Bayesian surprise [19], *i.e.*, which states things that are different on the temporal domain attract attention, but with time they get incorporated into our world model and become less relevant. We keep this factor in mind when designing the experiment, as people who are not used to interacting with a robot might direct their attention to it just because it is something new, rather than because of its actions.

It is also important to mention that in order to attract the user's attention, the robot has to be attentive to the person. This often involves mimicking human mechanisms that indicate attention. Bruce, Nourbakhsh and Simmons [4] found that if a robot turns its head to the person whose attention it wants, then the probability of the person cooperating with the robot is greatly increased. This is also exploited by Embgen *et al.* [20], who further concluded that the robot needs only move its head to transmit its emotional state. Nevertheless, no further analysis was performed to determine whether or how this robot-to-human non-verbal communication impacts HRI.

E. *Linear Regression Classification*

LRC is a simple yet powerful method for classification based on linear regression techniques. Naseem, Togneri and Bennamoun [21] introduced LRC to solve the problem of face identification by representing an image probe as a linear combination of class-specific image galleries. This is performed by determining the nearest subspace classification and solving the inverse problem to build a reconstructed image, choosing the class with the minimum reconstruction error. During the training phase, the inputs are added to the database using a greedy approach. Every input image is required to add a minimum information gain in terms of linear subspace independence: only if they fulfill the criterion, they are added to the database. This keeps the database size small, and allows for efficient training and classification. To the best of our knowledge, we are the first to apply this method to raw depth image data.

III. Problem Formulation

The problem that we address is human attention estimation from a first-person perspective. At a coarse level, we define attention in three categories: a) *null* attention, b) *partial* attention and c) *total* attention. We believe that this simple scale is enough to model a wide range of situations, since they encode the willingness of a user to engage in interaction. If robots are meant to be efficient social agents, it is imperative to be able to detect the right instance to start, maintain and end task-oriented interactions with humans.

A. *Scenario*

In our study, we assume that the robot wants to interact with a human who is sitting at a desk, a common occurrence in an office environment (see Figure 1); the robot wants to start an interaction approaching the human from one side. The situation is analog to a human approaching a coworker at her desk, willing to know if she is available for a given task. For our experiments, we use the Willlow Garage PR2 robot, with users occupying lab workstations with computer terminals. Having the experiment occur within the confines of the lab spaces ensures users' attentions are not unduly attracted to the robot, as having the robot around is a fairly common occurrence in the lab.

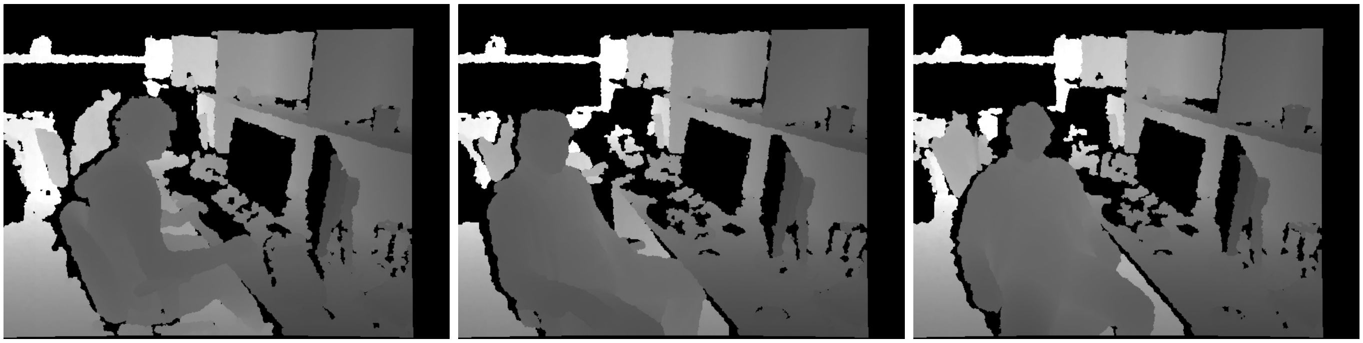


Figure 2. Representative raw-depth images of the three levels of attention. From left to right, *null* attention, *partial* attention and *full* attention. The images were captured using a Kinect$^{\text{TM}}$ mounted on the PR2 head.

B. Data

In order to evaluate human attention and obtain the best moment to ask for support, the robot relies only on a set of depth images captured with its Kinect$^{\text{TM}}$. For initial tests, the data was captured using a separate Kinect$^{\text{TM}}$ sensor mounted on a tripod simulating the pose that the sensor would have above the PR2. For our study, the data captured consisted only of depth information, allowing our approach to be robust against illumination variations, as well as other appearance changes. The intention is to demonstrate that our approach is robust enough so that visual information is not required, and efficient enough to run on a constrained computational platform while performing in real-time.

To build the training database, a subject is seated at a desk and asked to perform activities that simulate the three levels of attention of our scale. We describe each attention levels next:

1) *Null attention (class 1):* the subject's posture is such that she is facing the computer monitor, pretending she is busy, working;
2) *Partial attention (class 2):* the subject's posture is such that she is not facing the monitor, nor the robot, but rather facing somewhere in between;
3) *Full attention (class 3):* the subject's posture is such that she is facing the robot.

The subject is free to simulate the three attention levels according to her discretion, as long as the basic guidelines described above are satisfied. We recorded the movements of the subject and repeated the experiment several times; each one by placing the depth sensor in different configurations (*i.e.*, changing position, elevation and orientation), allowing for a more versatile training set. Examples of the range data are shown in Figure 2.

IV. Technical Approach

Our approach towards determining attention levels consists of an offline training stage and an online detection stage. The training step includes capturing depth snapshots (or video streams) of the user at her workstation or desk, extracting features and constructing class-specific feature matrices to build an attention classifier. During attention classification, instantaneous depth image snapshots from the Kinect$^{\text{TM}}$ are fed into the classifier, and the class with the minimum linear independence with respect to the training data is chosen as the likely attention level. The following sections provide technical details of these individual steps.

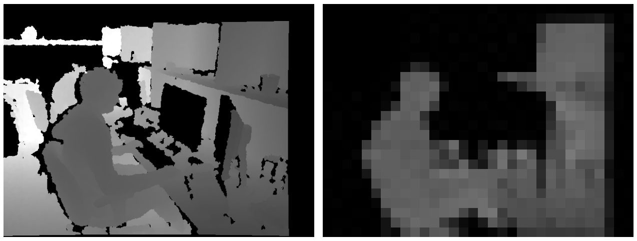


Figure 3. *Left*: Raw depth image and *Right*: preprocessed image ($\gamma = 1/20$ and distance = 2 meters).

A. Interaction Setting

The HRI is carried out as follows. First, the robot approaches and stands on either side of the human, using the range data to assess if the human is occupied, and if that is the case, continue evaluating the best moment to ask for support. If the human remains busy for an extended duration, then the robot does not engage in interaction and attends to other tasks. The aim is to evaluate if a robot standing close to a human working at a desk can accurately estimate the degree of attention of the human, and use this information to ask for support at the correct time instance. Data for our training set is obtained from the Kinect$^{\text{TM}}$ mounted on the robot in such scenarios, and is limited to range data only. Collected range images consist of particpants performing actions corresponding to the attention levels that we defined. The image sensor is placed on both sides of the user (see Figure 4), while recording the actions of the subject.

B. Features

For our LRC, the features consist of depth image data downsampled to $\gamma = \frac{1}{20}$ scale (see Figure 3), and reshaped by concatenation of its columns, similar to the methodology of Naseem, Togneri and Bennamoun [21]. However, as we are working with depth images, and we do not want the scene background to interfere with the learning and classification processes, the depth images are preprocessed to remove unwanted data. Specifically, we consider only those depth values up to a specific range, which accounts for the approximate physical distance between a robot and a human sitting at his desk under the current interaction scenario. This distance was empirically observed to be approximately 2 meters from the Kinect$^{\text{TM}}$ sensor.

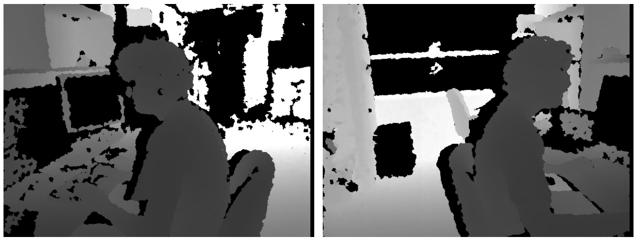


Figure 4. Range data from the left and right profile of the subjects. Both were included on the database for this study.

C. Training

Naseem, Togneri and Bennamoun [21] consider a database with photographs of people's faces. This is translated into a small number of sample images per class (subject), avoiding the time constraint of solving the pseudoinverse involved on a linear regression, experienced on large datasets, like videos. The novelty of our work is that we deal with this constraint (big datasets) by analyzing the linear independence of the images. By doing this, we dismiss all the new pictures that does not add relevant information to the LRC. Thus, we can condense the dataset into representative images, without losing relevant information. This allows us in turn to achieve an efficient LRC in real time.

Let $\mathrm{Y} \in \mathbb{R}^m$ be a vector of a given matrix $\mathrm{X} \in \mathbb{R}^{m \times n}$. We used a linear regression technique to analyze the linear independence of Y, as shown in Algorithm 3. In this algorithm, vector Y is projected onto the column space of X, then an error is calculated by subtracting the resultant projected vector Y_c to the original vector Y, giving as result the projection error ε. This ε is a metric that is used to measure the linear independence of Y with respect to the column space of X.

D. Algorithm

The overall algorithm is divided into two parts,

1) *Build X:* This procedure (Algorithm 1) analyzes the range image database and builds one matrix *per* attention class that contains the most significant collection of images in that class, and ensures a maximum degree of linear independence between images in the same class.
2) *Classify Y:* This procedure (Algorithm 4) is responsible for using the class matrices generated by the *Build X* algorithm, as well as the input depth image, to classify that image into one of the known classes. It also outputs the projection error of the image with respect to the column space generated by each of the class matrices, choosing the class with the minimum projection error.

It is important to mention that in order to reduce the overhead of a pseudo inverse calculation, $\mathrm{X}_i^\dagger$ is calculated only when X_i changes, thus X_i and $\mathrm{X}_i^\dagger$ are saved and computed only once for classification.

V. Experimental Results

We conducted a number of trials to evaluate our proposed approach. To train our system, we used range data of the three specific attention classes from 5 different participants, following the process described in Section IV. For each participant in the training process, we obtained video streams for three different attention levels, in two different Kinect$^{\mathrm{TM}}$ positions; each of the video streams have dimensions of 640×480 pixels, and have approximately 500 frames each. This resulted in a total of $15,000$ frames for the training process. The attention matrices X_i have average dimensions of 39×768, with 39 images downscaled to $\frac{1}{20}$ of their original dimensions. The value of the threshold τ was empirically set at 4.0 for all trials. Training and classification was performed on a PC with an Intel Core-i5$^{\mathrm{TM}}$ processor running at 1.7 GHz, with 2GB of memory and under the Ubuntu 12.04 Long-term Release (LTS) edition. The code was implemented in C++ using the Robot Operating System (ROS) C++ bindings.

Algorithm 1 Build X, the probe database.

```
Require: Threshold $\tau$
for each class $i$ do
    Img ← Random unseen image.                      ▷ Initialize $\mathrm{X}_i$
    Y ← FEATURES( Img )
    Append Y to $\mathrm{X}_i$
    Compute $\mathrm{X}_i^\dagger$                                  ▷ Build $\mathrm{X}_i$
    for each new image nImage do
        Y ← FEATURES(nImage)
        $\varepsilon$ ← LINEARINDEPENDENCE$(\mathrm{Y}, \mathrm{X}_i, \mathrm{X}_i^\dagger)$
        if $\varepsilon \geq \tau$ then
            Append Y to $\mathrm{X}_i$
            Compute $\mathrm{X}_i^\dagger$
        end if
    end for
end for
save($\mathrm{X}_i^\dagger$, X)
```

Algorithm 2 Feature extraction. Performs downsampling and reshaping.

```
procedure FEATURES( Image, $\gamma$ )
    Cut the image background.
    Down-sample the image by a factor $\gamma$.
    Reshape the image to a column vector.
    return The post-processed Image.
end procedure
```

Our results are summarized in Figures 5 and 6. The figures show frame-by-frame reconstruction errors of a test video. The errors represents the linear independence of an input image with respect to the column space of each class-specific database X_i, $i \in \{1, 2, 3\}$.

Figure 5(a) shows the classification of a video that was used during the training phase, as expected, the projection error is close to zero almost all the time. When the error reaches zero, is an indication that the current image passed the linear independence test, and it was used on the learning phase. While this is not illustrative of the accuracy of our algorithm, it does illustrate the fact that most of the information used during training is redundant, and that by keeping only a small fraction of it we can achieve a low reconstruction error. Figure 5(b) shows the classification performance on a video sequence that was not used during the training phase. While the reconstruction error is larger in this case, it is nonetheless sufficient to perform classification accurately. The key observation is that irrespective of the actual reprojection error numbers, there is clear separation between the actual attention class errors and the errors of the other attention classes, which leads to distinct identification of the user's attention level.

Algorithm 3 Measure the linear independence of Y with respect to database X.

```
Require: $Y \in \mathbb{R}^m$, $X, \in \mathbb{R}^{m \times n}$, $X^\dagger \in \mathbb{R}^{n \times m}$
procedure LINEARINDEPENDENCE(Y, X, $X^\dagger$)
    $Y_c \leftarrow XX^\dagger Y$                      ▷ Reprojection of Y.
    $\varepsilon \leftarrow ||Y_c - Y||^2$
    return $\varepsilon$                  ▷ The reprojection error is the metric.
end procedure
```

Algorithm 4 Classify a new vector Y.

```
Require: An input Image. Precomputed $X_i$ and $X_i^\dagger$ for each class $i$,
    the class-specific databases and their pseudo-inverses.
for each class $i$ do
    $Y \leftarrow$ FEATURES( Image )
    $\varepsilon_i \leftarrow$ LINEARINDEPENDENCE$(Y, X_i, X_i^\dagger)$
end for
return $\text{argmin}_i(\varepsilon_i)$        ▷ Return the class with minimum error.
```

The LRC is done *per* frame, by choosing the minimum of these errors, and using a leave-one-out validation process during training (*i.e.*, while building X, one subject was left out the matrix). Hence, as is observed in Figure 5, the robot is capable of correctly estimating the attention level, even when testing on a subject that was not included on the database. Figure 7 shows the difference in poses between the training (left) and testing (right) sets.

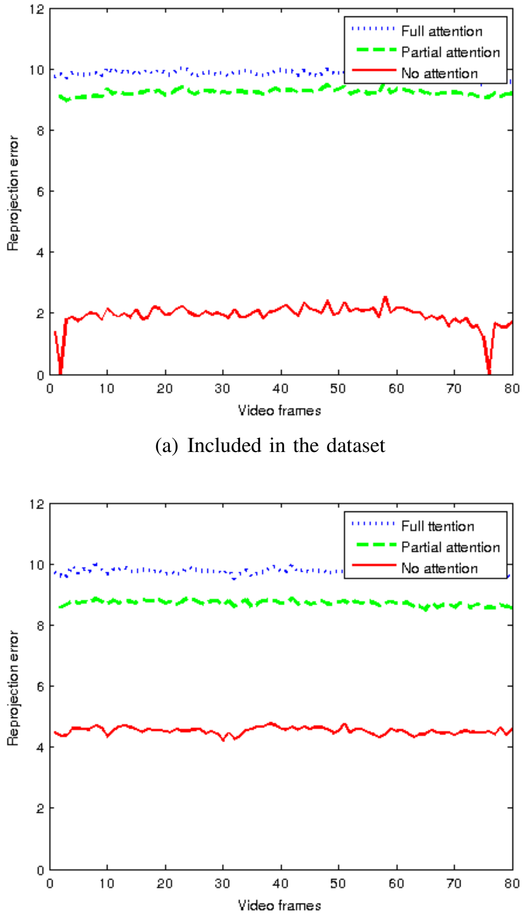
Nevertheless, large variations in the RGB+D sensor pose can lead to reduced performance of our algorithm. This is demonstrated in Figure 6, where the robot (and thus the Kinect$^{\text{TM}}$) is continually placed in positions not used to capture training data, while the system tries to detect a user in Class 3 (*i.e.*, full) attention level. As the Kinect$^{\text{TM}}$ changes its pose, the errors levels vary, resulting in an inaccurate classification. Note that the seperations between classes on the error scale are also reduced, resulting in degraded accuracy. The slopes on the figure represent displacement of the Kinect$^{\text{TM}}$.

A. Quantitative Analysis

In order to evaluate and validate our proposed method, we compare it against other common approaches for estimating visual attention based only on visual information, namely the Head Pose and Gaze attention estimation [22][14]. Ideally the comparison should be carried out using a ground truth of the subject attention, but this is extremely subjective, due to the inherently complexity of the human behavior, for this reason a simulated labeled attention is used as ground truth. To simulate this baseline, attention is lurked to the camera by showing interesting images to the user in a display beneath the sensor; similarly the attention is also directed outside the camera showing interesting images in another display.

For the purpose of this evaluation, the three attentive states are wrapped into two main states, attention or no attention towards the sensor, so, when the estimator and the simulated ground truth coincide on the attentive state of the user, a 1 or *OK* is recorded, otherwise a 0 or *NOT OK* is recorded. In the end, all this records are averaged in order to obtain an average accuracy of the estimator against the simulated ground truth, which in the context of this paper, it can be used as the estimator's main metric of performance.

The performance comparison is summarized on the Table I.



(a) Included in the dataset

(b) Unkown subject

Figure 5. Reprojection class error vs. Time. An input video from Class 1 (no attention) is being classified in real time. *(a)* LRC over a dataset that was included on the learning database. *(b)* LRC over data that was not included on the learning database.

TABLE I. Comparison of proposed Raw-Range-Information attention estimator with approaches based purely on visual information, Head Pose and Eye Gaze (coarse). Avg. Accuracy corresponds to the attention state between that reported by the estimator and the labeled attention.

Estimator	*Avg. Accuracy*
HeadPose-based	0.7333
EyeGaze-based	0.5667
LRC on raw range (proposed)	**0.8500**

VI. Conclusions and Future Work

We have presented a novel method to accurately estimate the attention of a person when interacting with a robot.

The results demonstrate that our method outperforms other models for attention estimation based solely on visual information (Table I).

The estimation is performed using raw depth data of the person's body pose. We use a LRC with the selection of the best linearly independent depth-images during the training phase. Our method also effectively discards redundant information during the training phase, while maintaining good performance on previously-unseen sequences. In addition to being completely independent from appearance and illumination changes, our approach is robust to small pose variations,

ISBN: 978-1-61208-390-2

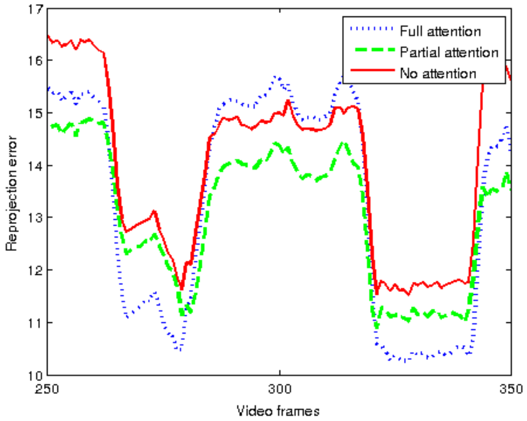


Figure 6. Effect of sensor displacement on classifier performance (Full attention).

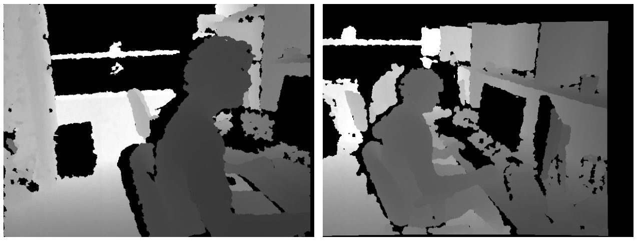


Figure 7. Difference in the position of the range sensor between images that were included (*left*) and not included (*right*) on the database for this study.

from the training data. The main advantages of our classifier are its simplicity, real-time computability and small memory footprint, which is ideal for implementation on board robots for man-machine interaction tasks.

In future work, we intend to explore attention estimation in a wider variety of settings, perform larger-scale experiments (encompassing more attention classes, more subjects, more situations), and explore the limits of the LRC approach. Particularly, we plan to fuse this approach with other face and gaze detectors in order to achieve a short-long distance attention estimator.

Acknowledgements

We would like to thank CONACyT and CBIE, under the ELAP, as well as Fonds de Recherche du Quebéc – Nature et Technologie, for their support and project funding.

References

[1] M. Argyle, Bodily communication. Routledge, 2013.

[2] M. Knapp, J. Hall, and T. Horgan, Nonverbal communication in human interaction. Cengage Learning, 2013.

[3] N. Hadjikhani, K. Kveraga, P. Naik, and S. P. Ahlfors, "Early (n170) activation of face-specific cortex by face-like objects," Neuroreport, vol. 20, no. 4, 2009, p. 403.

[4] A. Bruce, I. Nourbakhsh, and R. Simmons, "The role of expressiveness and attention in human-robot interaction," in Robotics and Automation, 2002. Proceedings. ICRA '02. IEEE International Conference on, vol. 4, 2002, pp. 4138–4142.

[5] S. Lang, M. Kleinehagenbrock, S. Hohenner, J. Fritsch, G. A. Fink, and G. Sagerer, "Providing the basis for human-robot-interaction: A multi-modal attention system for a mobile robot," in Proceedings of the 5th International Conference on Multimodal Interfaces, ser. ICMI '03. New York, NY, USA: ACM, 2003, pp. 28–35. [Online]. Available: http://doi.acm.org/10.1145/958432.958441

[6] A. Weiss, J. Igelsbock, M. Tscheligi, A. Bauer, K. Kuhnlenz, D. Wollherr, and M. Buss, "Robots asking for directions - the willingness of passers-by to support robots," in Human-Robot Interaction (HRI), 2010 5th ACM/IEEE International Conference on, March 2010, pp. 23–30.

[7] J. Shotton, A. Fitzgibbon, M. Cook, T. Sharp, M. Finocchio, R. Moore, A. Kipman, and A. Blake, "Real-time human pose recognition in parts from single depth images," in Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on, June 2011, pp. 1297–1304.

[8] Y. Yang and D. Ramanan, "Articulated pose estimation with flexible mixtures-of-parts," in Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on, June 2011, pp. 1385–1392.

[9] P. Felzenszwalb, D. McAllester, and D. Ramanan, "A discriminatively trained, multiscale, deformable part model," in Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on, June 2008, pp. 1–8.

[10] K. Fragkiadaki, H. Hu, and J. Shi, "Pose from flow and flow from pose," in Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on, June 2013, pp. 2059–2066.

[11] F. Kondori, S. Yousefi, H. Li, and S. Sonning, "3d head pose estimation using the kinect," in Wireless Communications and Signal Processing (WCSP), 2011 International Conference on, Nov 2011, pp. 1–4.

[12] E. Murphy-Chutorian and M. Trivedi, "Head pose estimation in computer vision: A survey," Pattern Analysis and Machine Intelligence, IEEE Transactions on, vol. 31, no. 4, April 2009, pp. 607–626.

[13] D. Hansen and Q. Ji, "In the eye of the beholder: A survey of models for eyes and gaze," Pattern Analysis and Machine Intelligence, IEEE Transactions on, vol. 32, no. 3, March 2010, pp. 478–500.

[14] A. Doshi and M. Trivedi, "Attention estimation by simultaneous observation of viewer and view," in Computer Vision and Pattern Recognition Workshops (CVPRW), 2010 IEEE Computer Society Conference on, June 2010, pp. 21–27.

[15] B. Benfold and I. Reid, "Guiding visual surveillance by tracking human attention," in Proceedings of the 20th British Machine Vision Conference, September 2009, pp. 1–11.

[16] M. Ryoo and L. Matthies, "First-person activity recognition: What are they doing to me?" in Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on, June 2013, pp. 2730–2737.

[17] P. Dollar, V. Rabaud, G. Cottrell, and S. Belongie, "Behavior recognition via sparse spatio-temporal features," in Visual Surveillance and Performance Evaluation of Tracking and Surveillance, 2005. 2nd Joint IEEE International Workshop on, Oct 2005, pp. 65–72.

[18] L. Itti and C. Koch, "Computational modelling of visual attention," Nature reviews neuroscience, vol. 2, no. 3, 2001, pp. 194–203.

[19] L. Itti and P. Baldi, "Bayesian surprise attracts human attention," Vision research, vol. 49, no. 10, 2009, pp. 1295–1306.

[20] S. Embgen, M. Luber, C. Becker-Asano, M. Ragni, V. Evers, and K. Arras, "Robot-specific social cues in emotional body language," in RO-MAN, 2012 IEEE, Sept 2012, pp. 1019–1025.

[21] I. Naseem, R. Togneri, and M. Bennamoun, "Linear regression for face recognition," Pattern Analysis and Machine Intelligence, IEEE Transactions on, vol. 32, no. 11, Nov 2010, pp. 2106–2112.

[22] A. Doshi and M. Trivedi, "Head and gaze dynamics in visual attention and context learning," in Computer Vision and Pattern Recognition Workshops, 2009. CVPR Workshops 2009. IEEE Computer Society Conference on, June 2009, pp. 77–84.