Evaluating Real-Time Mirroring of Head Gestures using Smart Glasses

Juan R. Terven
CICATA-IPN
Queretaro, Mexico
j.r.terven@ieee.org

Bogdan Raducanu
CVC
Bellaterra, Spain
bogdan@cvc.uab.es

María-Elena Meza
UAQ
Queretaro, Mexico
mezamariel@gmail.com

Joaquín Salas*
CICATA-IPN
Queretaro, Mexico
jsalasr@ipn.mx

Abstract

Mirroring occurs when one person tends to mimic the non-verbal communication of their counterparts. Even though mirroring is a complex phenomenon, in this study, we focus on the detection of head-nodding as a simple non-verbal communication cue due to its significance as a gesture displayed during social interactions. This paper introduces a computer vision-based method to detect mirroring through the analysis of head gestures using wearable cameras (smart glasses). In addition, we study how such a method can be used to explore perceived competence. The proposed method has been evaluated and the experiments demonstrate how static and wearable cameras seem to be equally effective to gather the information required for the analysis.

1. Introduction

Mirroring occurs when one person mimics the non-verbal communication (head movements, hand gestures, facial expressions, tone of voice, verbal accent, breathing, etc.) of their counterparts. The role of mirroring is to signal empathy between people and is an early indicator of a positive outcome (agreement or recognition) in an interaction. It has been observed that head gestures, such as nodding, increase the opportunities for a person to be liked [6], while the occurrence of behavioral mirroring is an early predictor of acceptance [11, 18, 19]. For some applications, as in negotiation, it is essential to have real-time feedback assessing the state of the conversation, a prediction of the interaction outcome, or whether there has been a break-point between the interlocutors.

Even though mirroring is a complex phenomenon [5, 7, 30], in this study we focus on the detection of head-nodding as a simple non-verbal communication cue due to its significance as a gesture displayed during social interactions. Apart from the obvious function of signaling a *yes*, head nods are used as backchannels to display interest, enhance communication or anticipate the counterpart's intention for turn claiming [1, 25]. In the current work, head noddings are recognized using cumulative histograms of facial features (as extracted by the algorithm introduced by Xiong and De la Torre [34]). After nodding recognition, mirroring is inferred based on a variable temporal window between two sequences of noddings. As our results reflect, even this reduced perspective of the problem is useful to robustly infer mirroring using either a static or a wearable camera.

This paper introduces a computer-vision method to detect mirroring through the analysis of head gestures using wearable devices. In our scenario, there is a dyadic conversation taking place between a service provider and a customer (see Figure 1). The social interaction is recorded with wearable cameras (see the smart glasses in Figure 2), and a pair of fixed-cameras, facing each participant. The smart glasses possess a high-definition video-camera located in the bridge between the two lenses. The role of the fixed cameras is twofold: (i) To compare the performance of the wearable devices against stable video and (ii) to extract ground-truth for head nodding detection. These recordings generate a set of images designated by $\mathcal{I}(t)$. These images are fed into a head-gesture recognizer for automatic nodding recognition $\eta(t)$. The synchronous recognition of head nods leads to mirroring detection $\mu(t)$. After each recorded session, a set of qualitative interviews was carried out with the *customers* in order to assess their perception on the competence of the service provider, thus creating the classification space $\mathcal{C}$. The analysis of these interviews allowed us to perform a competence assessment $\mathcal{L}$ (*i.e.*, if the customer considered that the service provider was competent in addressing her/his requests). Our goal was to see if we could reach a similar result by automatic detection of the mirroring activity.

Our main contribution is twofold: (i) We propose a solution for automatic mirroring detection based on wearable technology, *i.e.*, smart glasses, and (ii) the proposed algorithm is used for competence assessment, by analyzing the

*Corresponding author: Joaquín Salas. Instituto Politécnico Nacional. Cerro Blanco 141, Colinas del Cimatario, Querétaro, México, 76090. jsalasr@ipn.mx.

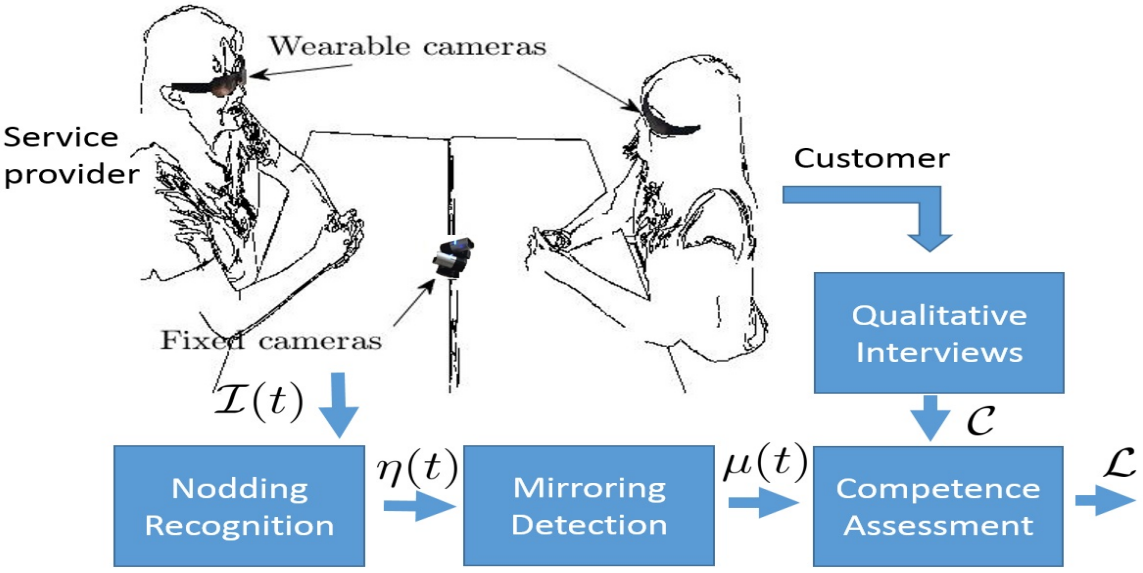


Figure 1. Visual representation of the steps involved in the automatic mirroring detection and its application to competence assessment

amount of mirroring activity generated during a dyadic conversation.

The choice for using wearable cameras is motivated by the fact that such devices offer a first-person perspective, compared to the classical fixed cameras, which offer a third-party perspective. For this reason, we strongly believe that the results of our work will motivate the use of smart glasses as assistive technology for people suffering of a wide range of disorders of the visual system: Visual impairment, prosopagnosia (face blindness), Asperger syndrome (difficulty to decode/understand emotions and other social signals), etc. Furthermore, the compact design of this device (due to the fact the camera is embedded in an everyday artifact used by a significant number of people), is a guarantee for a high probability of its acceptance.

The rest of the paper is organized as follows. In Section 2, we provide an overview of the related literature on automatic recognition of mirroring behavior. Section 3 presents our approach on automatic mirroring recognition. In Section 4, we present the experimental results for mirroring detection and its relation to competence assessment. Finally, in section 4.3 we present our conclusions and provide guidelines for future work.

2. Related Work

Although psychological research on role analysis dates back to the early 1970s, the computational approach for studying mirroring behavior has only recently started to address this problem. Several technologies have been used, but the one represented by computer vision occupies a central role. Some comprehensive surveys on this topic can be found in Delaherche *et al.* [9] and Wagner *et al.* [32]. In the

Figure 2. Smart glasses used as wearable camera. The glasses have a high-definition camera embedded in the bridge connecting the two lenses (by Pivothead Inc., with permission).

remaining of this section, we review only the most relevant works related to our research.

Ramseyer and Tschacher [26] estimated mirroring based on the cross-correlation of motion energy features computed over a temporal window of a few seconds. Here, motion energy is defined as the difference between consecutive frames and is used as a measure of global value of activity. Their experiments demonstrated that nonverbal synchrony is higher in genuine interactions when contrasted with pseudo-interactions. A similar approach has been reported by Sun *et al.*[28] with a more sophisticated analysis. In their approach, the information provided by motion energy is converted into histograms, but the image is not processed holistically. Instead, the region containing the body parts is divided in sub-regions through quad-tree decomposition for more efficient feature extraction. Subsequently, they use a traditional template-based action recognition approach to compute behavior similarities of corresponding temporal windows (sequence of frames). This

approach has been tested upon a custom dataset containing people engaged in face-to-face conversation.

These techniques have been replaced by more refined computer vision techniques. For instance, Girard *et al.*[16] showed the feasibility of measuring automatically facial expressions included in the Facial Action Coding System (FACS)[10]. In their experiment, they analyze 400,000 video frames, obtained from fixed cameras, from 80 people. For automatic detection of facial expression, they automatically extracted facial landmarks[34], measured the face deformation with SIFT features, and trained SVM classifiers to detect each expression. Their results show that fewer than 6% of the frames had to be coded manually. In our case, we also extract facial landmarks but as a step toward the construction of 3D models and head pose estimation. The automatic measures of facial actions and head motion have been used as behavioral predictors[15].

The previous approach has been extended by also incorporating the non-verbal information contained in the audio channel. More precisely, Sun *et al.* [29] extracted from the acoustic signal the following prosodic features: Pitch, intensity, energy and speaking rate. Following a similar multimodal framework, Delaherche and Chetouani [8] studied synchrony on a cooperative task where both partners have to coordinate in order to build an assembled object. To identify the coordination between demonstrator and experimenter, they used the Pearson correlation and magnitude coherence between all pairs of features. For instance, they found that the lowest percentage of coordination was obtained in pitch and pause. On the other hand, regarding the visual domain, it is apparent that the image of motion history is the feature that best captures the synchrony of actions. Also in a multimodal framework, Bilakhia *et al.*[4] proposed a method to detect mimicry behavior in audiovisual data. They used a corpus of naturalistic dyadic interactions, and their approach was based on a temporal regression model, represented by long short-term memory networks, in order to reproduce one subject's behavior from the other's. Recently, Bilakhia *et al.*[3] introduced `MAHNOB`, an annotated set of audiovisual recording of dyadic interactions for the study of mimicry behavior. Their video is obtained with fixed cameras, which made it unsuitable for our research purposes. As we based our study on head-nods, it is interesting to note that about 60% of the 11 gestures they annotate are noddings. In their experiments, they consider eleven face and head movements, which include smile, laughter, frown, eyebrow raise, head nod, head shake, head pose shift, shoulder shrung, left hand movement, right hand movement, and torso shift.

Recently, Michelet *et al.* [22] presented an unsupervised method to estimate mirroring. For this purpose, they computed *Bag-of-Words* models [13] around some feature points from the spatio-temporal analysis of the sequence. Then, similarity between bag-of-words models is measured with dynamic-time-warping, giving an accurate measure of imitation between partners. A threshold has been used in order to discriminate between mimicry and non-mimicry. A similar approach, based on time-series processing, has also been pursued [27, 24]. At their end, Messinger *et al.* [21] studied the face-to-face interaction between infants and their parents. More concretely, they introduced a machine learning framework to explore the predictability of infant-mother behavior. For instance, it was expected that mothers smiled predictably in response to the smiles of the infants, and the initiation of the smiles of the infants become more predictable over developmental time. The smiles have been manually annotated in video data using FACS [10]. Two types of models have been used: A causal and a temporal one which were characterized in terms of turn-takings. A turn-taking event was defined as a mother or infant transition that was immediately preceded by the transition of the other partner. Another interesting application was introduced by Yu *et al.*[35]. They develop an automated method to detect deception out of interactional synchrony. They used Face-to-Face interactions and video-conference sessions. In all their sessions, they used fixed cameras. They automatically track head gestures and expressions. They built an SVM classifier, which has a detection accuracy of 74%, compared to the 54% accuracy one could expect from a typical unaided individual. The characteristics they use include facial expressions, like smiling and looking forward, and head movements, like nodding. The most effective features, *i.e.*, the ones which achieves the best performance in the classification task, are selected via a Genetics algorithms. In our case, we have applied mirroring to the assessment of perceived competence.

In a nutshell, our research distinguish itself from the state-of-the-art in that we pioneer the use of wearable devices to detect mirroring and study its use in novel applications, as it is assessment of perceived competence. Overall, we contribute the body of research on that is being called social signal processing[23].

3. Mirroring Inference

The procedure for automatic mirroring detection consists of the following steps: 1) facial features extraction and stabilization; 2) head nodding gestures recognition; and 3) mirroring inference. A detailed explanation is provided for each of these steps in the remainder of this section.

3.1. Facial Features Extraction and Stabilization

To extract the facial features, we used non-rigid face tracking with the Supervised Descent Method (SDM) from [34]. Then, we applied a stabilization step to compensate for camera motion (user's ego-motion). Unlike traditional video stabilization approaches where the whole frame

is warped and smoothed in order to create a stable version of the video [2, 17, 20], in our case we only stabilize the tracked facial features. Figure 3 shows a comparison of the facial features position from a static camera, a wearable camera, and the stabilized version, respectively. The three plots display the facial features changes in the horizontal and vertical directions (orange and blue respectively). The red vertical lines mark the start and end of head gestures (nods). The middle plot shows the effect of the camera's ego-motion. This motion triggers false head gestures detection. The third graph shows the results of ego-motion stabilization. The stabilization process attenuates camera motion, preserving head motion for gesture recognition.

To stabilize ego-motion, our procedure detects and tracks background features; it then fits a motion model for the camera and finally, it compensates for camera motion. To estimate camera motion, we extracted sparse features of the background and track them using optical flow. We discarded features inside the face region to prevent the use of head motion as background motion. More formally, we created a set of n pairs of matched features, located in (x_{t-1}^j, y_{t-1}^j) and (x_t^j, y_t^j), for $j = 1, \ldots, n$, from the previous and current frames. Using this set of features, we estimated the interframe motion represented as a two-dimensional linear model with four parameters similar to the one proposed in Battiato *et al.* [2]:

$$\begin{bmatrix} x_t^j \\ y_t^j \\ 1 \end{bmatrix} = \begin{bmatrix} \lambda\cos\phi & -\lambda\sin\phi & T_x \\ \lambda\sin\phi & \lambda\cos\phi & T_y \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} x_{t-1}^j \\ y_{t-1}^j \\ 1 \end{bmatrix},$$

$$\mathbf{x}_t^j = \mathbf{C}\mathbf{x}_{t-1}^j, \tag{1}$$

where ϕ is the rotation angle, T_x and T_y the translation in the x and y direction, and λ is a scale parameter.

Some tracked features may have different motions due to wrong matches or moving objects in the background. To remove these outliers from the set of features, we used localized RANSAC as described by Grundmann *et al.* [17]. Once we removed the outliers, we remained with a set of k feature pair and solved (1) for the four variables using linear Least Squares to obtain the motion matrix $\mathbf{C}$. Therefore, one can establish a relationship between the facial features with camera motion $\boldsymbol{x}^w = [x^w, y^w, 1]^T$ and the motionless facial features $\boldsymbol{x}^s = [x^s, y^s, 1]^T$ via the camera motion $\mathbf{C}$ as

$$\boldsymbol{x}^w = \mathbf{C}\boldsymbol{x}^s. \tag{2}$$

3.2. Head Nodding Gestures Recognition

Once we have stabilized the head motion, we created histograms of orientations (HOO) that will be used as predictors for a Random Forests classifier.

Raw motion observations are susceptible to noise and scale changes. Figures. 4(a) and 4(c) show the raw motion of a nodding gesture and other head gesture respectively. These motion time-series are noisy and can change dramatically at different distances and resolutions. To obtain distance and resolution invariant, we scaled down the motion using the face size (from the facial features). Then we extracted a compact representation of the motion using a Histogram of Orientations [14] with the following procedure (see algorithm 1): First we tracked a facial feature (*i.e.*, tip of the nose) during k consecutive frames, n and $n-1$, obtaining its motion $\mathbf{p} = (p_x, p_y)$, where $\mathbf{p} = \mathbf{p}_n - \mathbf{p}_{n-1}$; then we calculated its orientation, $\theta = \arctan(p_y, p_x)$ (taking care of the case $p_x = 0$, and magnitude, $M = \sqrt{p_x^2 + p_y^2}$; then we added the magnitude M to the bin $b = \lfloor \theta/360 \rfloor B$, which belongs to the B-bins histogram $\mathbf{h}$.

Call : $< \mathbf{h}, \mathbf{p}_n > \leftarrow$ HOO $(\mathbf{p}_n, \mathbf{p}_{n-1}, \mathbf{h})$
input : Corresponding feature $\mathbf{p}_{n-1}$ and $\mathbf{p}_n$, and histogram of orientations $\mathbf{h}$ with B bins
output: An updated histogram of orientations $\mathbf{h}$

// Compute the motion of a feature between frames $n-1$ and n. Here $\mathbf{p} = (p_x, p_y)$.
$\mathbf{p} \leftarrow \mathbf{p}_n - \mathbf{p}_{n-1}$;
// Calculate its orientation in the range $0° \leq \theta \leq 360°$. Take care of $p_x = 0$
$\theta \leftarrow \arctan(p_y, p_x)$;
// Calculate its magnitude
$M \leftarrow \sqrt{p_x^2 + p_y^2}$;
// Compute the histogram index
$b \leftarrow \lfloor \theta/360 \rfloor B$;
// Add the magnitude to the bin
$\mathbf{h}[b] \leftarrow \mathbf{h}[b] + M$;

Algorithm 1: Histogram of Orientations. The histogram of orientations $\mathbf{h}$ was computed across the frames of the sequence with the use of a particular feature. In the process, the bin B corresponding to a certain orientation θ was weighted by the magnitude of the movement M.

Figures. 4(b) and 4(d) show the non-normalized histogram for the corresponding gesture at its left. Notice that a nodding gesture, which is described as moving the head up and down has large bins at 90° and at 270°. On the contrary, other head gestures have contributions elsewhere without a concrete pattern.

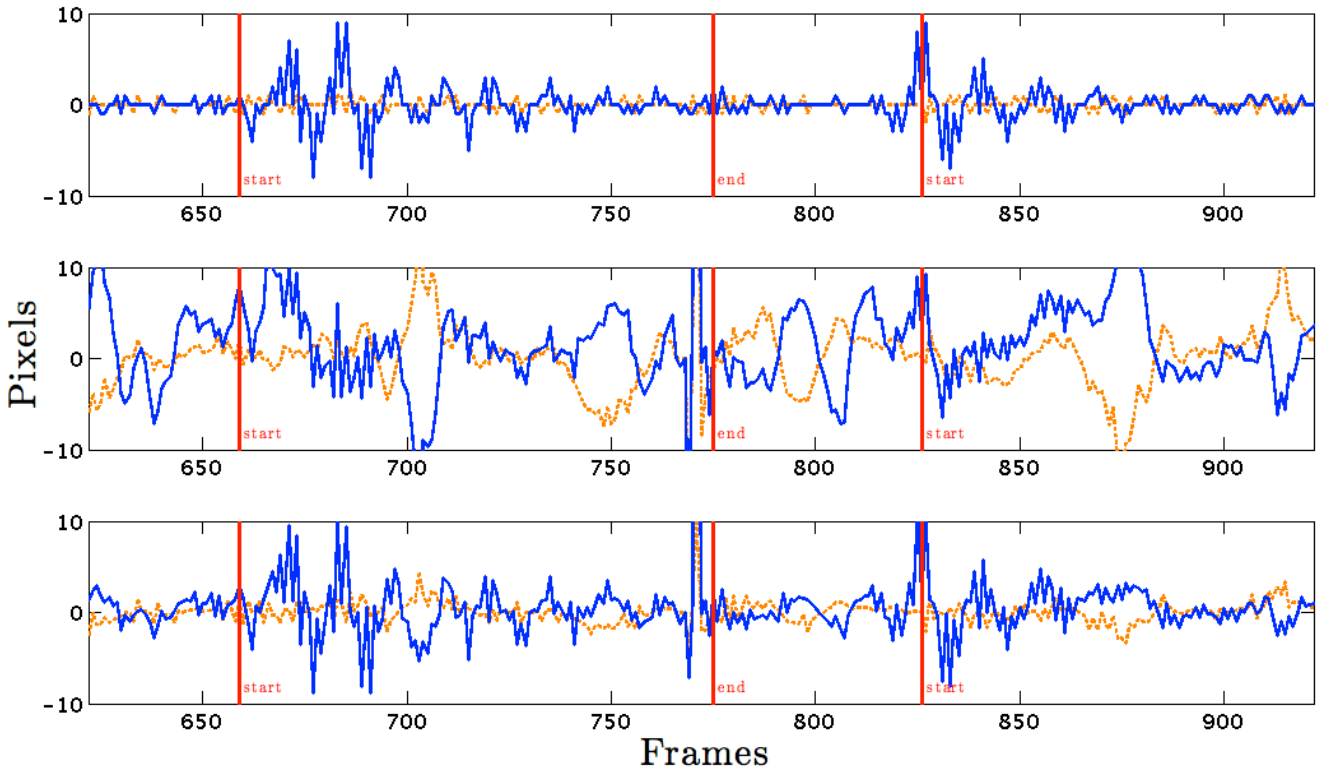


Figure 3. Head Motion. Horizontal and vertical motion (dotted and solid lines respectively) of the head in ten seconds of video. Vertical lines mark the start and end of head nodding. The first plot shows the motion inferred from a static camera video. The second plot shows the same event as viewed from a wearable camera, displaying the camera motion effect. The third plot shows the stabilized sequence.

3.3. Automatic Mirroring Detection

Inspired by the approach of Feese *et al.*[12], but measuring mirroring in both directions, we defined two events: *Person $\mathcal{A}$ is mirroring Person $\mathcal{B}$* or ($m\mathcal{AB}$); and *Person $\mathcal{B}$ is mirroring Person $\mathcal{A}$* or ($m\mathcal{BA}$). To count an $m\mathcal{AB}$ event, person $\mathcal{A}$ needs to start displaying gesture ξ after person $\mathcal{B}$ started and within a time Δt after person $\mathcal{B}$ stopped displaying gesture ξ. In case that person $\mathcal{A}$ displays ξ multiple times while $\mathcal{B}$ is displaying ξ, only one event is counted. Similarly, a $m\mathcal{BA}$ event is triggered when person $\mathcal{B}$ starts displaying gesture ξ after person $\mathcal{A}$ started and within Δt after person $\mathcal{A}$ stopped displaying gesture ξ. Gesture repetitions were treated the same way. More formally, given a sequence of gestures $g^{\xi}_{1\ldots N^{\xi}}$ of person $\mathcal{A}$, the start and end times of each gesture is given by $t_1\left(g^{\xi}_i\right)$ and $t_2\left(g^{\xi}_i\right)$ respectively. An $m\mathcal{AB}$ event is triggered if (following [12]):

$$\begin{gathered} g^{\mathcal{A}}_i = g^{\mathcal{B}}_j, \\ t_1\left(g^{\mathcal{B}}_j\right) < t_1\left(g^{\mathcal{A}}_i\right) < t_2\left(g^{\mathcal{B}}_j\right) + \Delta t. \end{gathered} \tag{3}$$

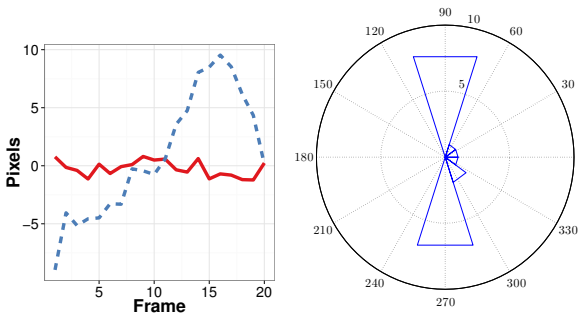
Figure 5 shows a fragment of 1,000 frames (about 30 seconds) from one of the videos in our dataset. The top row shows the nodding gestures performed by person $\mathcal{A}$, the middle row shows the nodding gestures performed by person $\mathcal{B}$, and the bottom row shows the mirroring events. The first three mirroring events are triggered by person $\mathcal{A}$. In the first event of this sequence, person $\mathcal{A}$ mirrors person $\mathcal{B}$ after person $\mathcal{B}$ stopped displaying the nodding gesture, but within a predefined window Δt. The other mirror events occur just after person $\mathcal{B}$ started the nodding gesture. The fourth mirroring event is due to the person's $\mathcal{B}$ response to the nodding gesture triggered by $\mathcal{A}$. The window Δt is heuristically determined, taken into consideration the analysis of our dataset, where the average time lapse between gestures is 1.36s.

4. Experimental Results

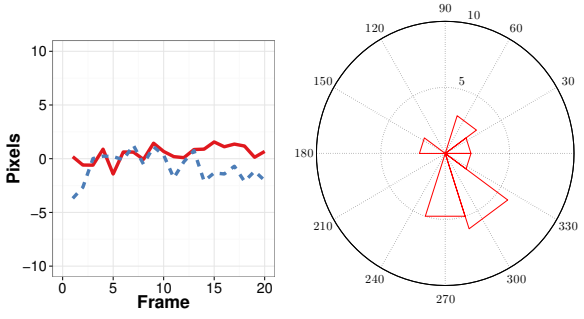
In this section we first describe our dataset. Then, we present the performance of automatic nodding recognition and mirroring detection. Finally, we perform a competence assessment based on the individual interviews conducted with each subject participating in the experiment.

4.1. Mirroring dataset

We defined a realistic conversation scenario, in which a student was instructed to ask a psychologist for advice regarding academic orientation and support. The conversations revolved around three questions: i) What to do in a



(a) Nodding motion (b) Nodding Histogram



(c) Other gesture motion (d) Other gesture Histogram

Figure 4. Head Gesture Motion and Histograms of Orientations (HOO) (best seen in color). (a) Shows the horizontal (solid red) and vertical motion (dashed blue) for a typical nodding gesture. (b) Shows the HOO of a nodding gesture, (c) shows the motion (solid red for the horizontal, and dashed blue for the vertical displacement) for an exemplary gesture other than nodding, and (d) shows its HOO.

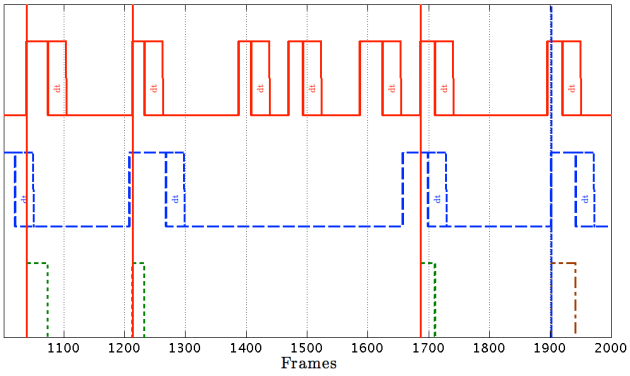


Figure 5. Mirroring detection (best seen in color). Rows one (solid red) and two (dashed blue) show the number of frames when gestures from either person $\mathcal{A}$ or person $\mathcal{B}$ occur. Note that there is a fixed interval of time dt when the mirroring effect may take place. The third row displays the occurrence of mirroring. In frames 1,050, 1,210, and 1,690 person $\mathcal{A}$ mirrored person $\mathcal{B}$ (green). In frame 1,900, person $\mathcal{B}$ mirrored person $\mathcal{A}$ (brown).

given situation? ii) What is the psychologist's experience with similar problems? and, iii) How many therapy sessions are needed? The confederated psychologist answered the questions to each participant in the same verbal style but controlling his nodding gestures. We created a mirroring dataset consisting of 48 sessions (with a unique student per session) of three minutes each, on average. We recorded each session with two static HD cameras and two wearable cameras. For the static cameras we used Microsoft LifeCam Studio, fixed on the table and aimed at each participant.

Three trained sociology students annotated the starting and ending times of the nodding gestures in the videos using ELAN Linguistic Annotator [33]. These annotations served as gestures' ground truth and we used them to calculate the mirroring ground truth following the approach described in Section 3.

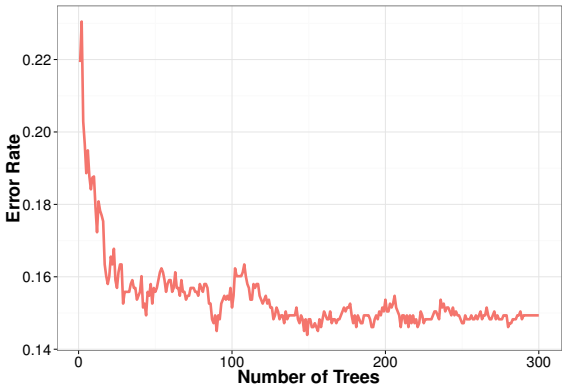


Figure 6. Training Error rate. This plot shows how the error rate changes with the number of trees during the training phase. In this plot we see that beyond 60 trees the error rate stays unchanged.

4.2. Head Nodding Recognition and Mirroring Detection

Since mirroring detection is mainly determined by correct head nodding recognition, we first report experimental results on it. Combining our dataset with the one presented in [31], we split our mirroring dataset into a training set (50%), a validation set (25%), and a test set (25%). From these sets, we extracted the histograms of orientations (HOO) and we trained a Random Forests classifier using the training set for nodding recognition. Figure 6 shows how the error rate varies with the number of trees during the training phase. From this plot we could appreciate that when the number of trees is above 60 the error rate remains stable. For this reason, to perform the remaining experiments, we kept this number of trees (60).

Another aspect we wanted to evaluate is how nodding duration affects the accuracy of the classifier, since the gestures in our dataset are variable in length, ranging from 15 frames (0.5s) up to 120 frames (4s). To evaluate the effect, we trained multiple Random Forests classifiers with variable length using our validation set. As result, we noticed

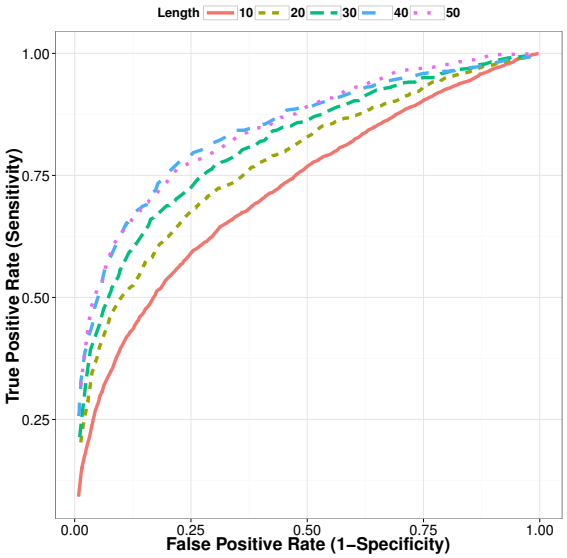


Figure 7. ROC Curves. This plot shows ROC curves for multiple gesture lengths ranging from 10 frames up to 50 frames. The curves show that the performance improves with larger gestures settling down around 40-50 frames.

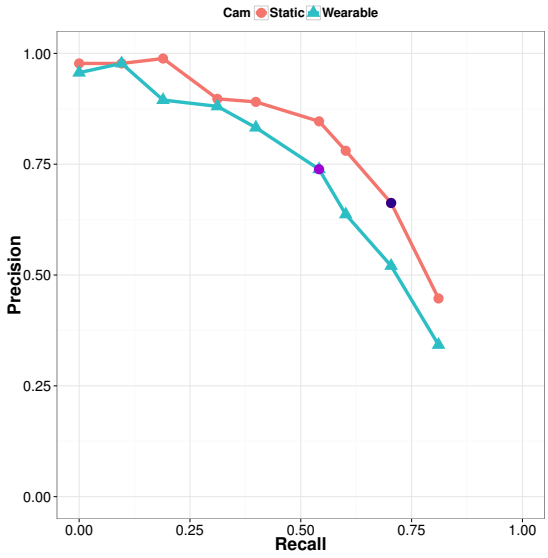


Figure 8. Mirroring Detection. The solid circle in the curves shows the position for the highest F_1-score in the validation set. For static cameras it is 0.68, while for wearable cameras it is 0.62.

that the recognition improves with increasing duration of gestures, stabilizing around 50 frames. Figure 7 shows a ROC curve for multiple gesture lengths ranging from 10 frames up to 50 frames.

Once we found the optimal value for parameters that guarantees an optimal head nodding recognition, we continued with the evaluation of mirroring detection using wearable cameras. We considered the results obtained with the same algorithm running on the fixed cameras as baseline. Using the validation set, we ran the mirroring detector multiple times varying the sensibility of the classifier. For this we used the confidence of the classification —calculated as the proportion of decision trees that classified the sample to the positive class— by varying the classification threshold between 0.1 to 0.9. For each case, we calculated the true positives (tp), false positives (fp), and false negatives (fn). With these values we constructed the Precision-Recall curve presented in Figure 8. Finally, we selected the model with the highest F_1-score given by

$$F_1 = \frac{2tp}{2tp + fp + fn}. \quad (4)$$

For the fixed cameras the F_1-score is 0.69, while for the wearable cameras it is 0.6. Figure 8 shows the position for the highest F_1-score as solid circles. In order to evaluate the generalization capability of our algorithm, we ran our classifier on the test set obtaining the results shown in Table 1. The reduction in performance of the wearable cameras is due to an increment of false positives given the camera motion.

Table 1. Mirroring performance on test set.

	Static camera	Wearable camera
Precision	0.80	0.65
Recall	0.62	0.62
F-Score	0.70	0.63

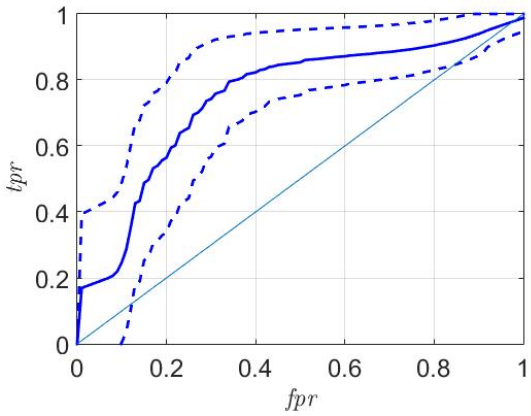
For experiments, our algorithm was implemented in C++ and was capable of detecting gestures at 20 frames/sec on HD video. The performance was achieved on a PC equipped with 8GB of RAM and an Intel Core i5 microprocessor running at 1.9GHz.

4.3. Competence Assessment

To demonstrate the usefulness of our algorithm for mirroring detection in the case of wearable devices, we apply it to assess the competence [1] derived from the dyadic social interactions. For this purpose we considered all four possible combinations of image sources: student←static camera, psychologist←static camera, student←static camera, psychologist←wearable camera, student←wearable camera, psychologist←static camera and student←wearable camera, psychologist←wearable camera. The ← symbol can be read as *was observed by the*.

As starting point for our study, we relied on the qualitative interview outcome that followed each session and where the student was asked about the satisfaction of the interaction. Based on the qualitative analysis of the inter-

[1] By competence assessment we refer to the ability of the psychologist to satisfactorily answer the student's questions - thus reflecting the student's viewpoint.



(a)

Figure 9. Perceived Competence. The ROC curve shows the average performance (solid line) for 1000 cycles of traning/testing, bounded by a standard deviation (dashed lines). *fpr*, *tpr* and *auc* stand for false positive rate, true positive rate, and area under the curve, respectively. Both the student and the psychologist were using wearable cameras. The area under the curve for the mean performance was 0.74.

views, we labeled the conversations contained in the *mirroring dataset* in two classes: competent or non-competent, based on the student's assessment. Using the number of times either the student or the psychologist mirrored each other, and the combination of images taken with the static and wearable cameras, we constructed a linear SVM classifier to distinguish between these two classes. Ideally, the SVM computes the hyperplane that best separates between both classes. However, due to the non-linearity between the two classes (and to minimize the number of outliers), we translated the hyperplane in order to obtain different classification performances. In each trial we used 50% of the interviews to construct the classifier and 50% to test it. To assess uncertainty we repeated this procedure 1000 times. From these 1000 trials we estimated the average and standard deviation for the performance. Figure 9 shows the ROC curve for the case where the student and the psychologist were using wearable cameras. As overall measure of performance we computed the area under the curve (auc), which for the mean performance gives 0.75, 0.74, 0.75 and 0.74, for the cases discussed above. Therefore, our results indicate that all combinations of static/wearable cameras seem to be equally supportive to evaluate the perceived competence.

Conclusion and Future Work

In this paper we presented a computer-vision based solution for automatic mirroring detection using wearable devices (i.e. smart glasses). Mirroring was detected by the analysis of a temporal window between two consecutive head noddings. We tested our approach on a custom dataset consisting of 48 dyadic conversations between a customer and a service provider. After that, we demonstrated that our approach can be successfully applied for competence assessment derived from the analysis of qualitative interviews recorded after each session. The encouraging results obtained so far motivate us to pursue the use of our approach in a real-world application. Future work will be devoted to enhance the dictionary of recognized head and facial gestures (shaking, emotional expressions, etc.) but also to explore its use as assistive technology for visually impaired people.

5. Ackowledgements

B. Raducanu is supported by the projects TIN2013-41751 and TIN2013-49982-EXP, MINECO, Spain. Juan Ramón Terven was partially supported by Tecnológico Nacional de México and CONACYT. Joaquín Salas was partially supported by FOMIX GDF-CONACYT under Grant No.189005, by IPN-SIP under Grant No. 20150281, and by UCMexus.

References

[1] J. Allwood and L. Cerrato. A Study of Gestural Feedback Expressions. In *Nordic Symposium on Multimodal Communication*, pages 7–22, 2003. 1

[2] S. Battiato, G. Gallo, G. Puglisi, and S. Scellato. SIFT Features Tracking for Video Stabilization. In *International Conference on Image Analysis and Processing*, pages 825–830, 2007. 4

[3] S. Bilakhia, S. Petridis, A. Nijholt, and M. Pantic. The MAHNOB Mimicry Database-a database of naturalistic human interactions. *Pattern Recognition Letters*, 2015. 3

[4] S. Bilakhia, S. Petridis, and M. Pantic. Audiovisual Detection of Behavioural Mimicry. In *International Conference on Affective Computing and Intelligent Interaction*, pages 123–128, 2013. 3

[5] E. Carr and P. Winkielman. When Mirroring is both Simple and *Smart*: How Mimicry can be Embodied, Adaptive, and Non-Representational. *Frontiers in Human Neuroscience*, 8:1–7, 2014. 1

[6] C. Catmur, V. Walsh, and C. Heyes. Associative Sequence Learning: The Role of Experience in the Development of Imitation and the Mirror System. *Philosophical Transactions of the Royal Society of London B: Biological Sciences*, 364(1528):2369—2380, 2009. 1

[7] T. Chartrand and J. Bargh. The Chameleon Effect: The Perception-Behavior Link and Social Interaction. *Journal of Personality and Social Psychology*, 76(6):893–910, 1999. 1

[8] E. Delaherche and M. Chetouani. Multimodal Coordination: Exploring Relevant Features and Measures. In *International Workshop on Social Signal Processing*, pages 47–52, 2010. 3

[9] E. Delaherche, M. Chetouani, A. Mahdhaoui, C. Saint-Georges, S. Viaux, and D. Cohen. Interpersonal Synchrony: A Survey of Evaluation Methods across Disciplines. *IEEE Transactions on Affective Computing*, 3(3):349–365, 2012. 2

[10] P. Ekman and W. Friesen. *The Facial Action Coding System: A Technique for The Measurement of Facial Movement*. Consulting Psychologists Press, 1978. 3

[11] S. Farley. Nonverbal Reactions to an Attractive Stranger: The Role of Mimicry in Communicating Preferred Social Distance. Journal of Nonverbal Behavior. *Journal of Nonverbal Behavior*, 38(2):195–208, 2014. 1

[12] S. Feese, B. Arnrich, G. Troster, B. Meyer, and K. Jonas. Quantifying Behavioral Mimicry by Automatic Detection of Nonverbal Cues from Body Motion. In *IEEE International Conference on Social Computing*, pages 520–525, 2012. 4, 5

[13] L. Fei-fei and P. Perona. A Bayesian Hierarchical Model for Learning Natural Scene Categories. In *Computer Vision and Pattern Recognition*, volume 2, pages 524–531, 2005. 3

[14] W. T. Freeman and M. Roth. Orientation Histograms for Hand Gesture Recognition. In *International workshop on automatic face and gesture recognition*, volume 12, pages 296–301, 1995. 4

[15] J. M. Girard and J. F. Cohn. Automated audiovisual depression analysis. *Current Opinion in Psychology*, 2014. 3

[16] J. M. Girard, J. F. Cohn, L. A. Jeni, M. A. Sayette, and F. De la Torre. Spontaneous Facial Expression in Unscripted Social Interactions can be Measured Automatically. *Behavior research methods*, pages 1–12, 2014. 3

[17] M. Grundmann, V. Kwatra, and I. Essa. Auto-Directed Video Stabilization with Robust L_1 Optimal Camera Paths. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 225–232, 2011. 4

[18] N. Guéguen, C. Jacob, and A. Martin. Mimicry in Social Interaction: Its Effect on Human Judgement and Behavior. *European Journal of Social Sciences*, 8(2):253–259, 2009. 1

[19] C. Jacob, N. Guéguen, A. Martin, and G. Boulbry. Retail Salespeople's Mimicry of Customers: Effects on Consumer Behavior. *Journal of Retailing and Consumer Services*, 18(5):381–388, 2011. 1

[20] Y. Matsushita, E. Ofek, W. Ge, X. Tang, and H.-Y. Shum. Full-Frame Video Stabilization with Motion Inpainting. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28(7):1150–1163, 2006. 4

[21] D. Messinger, P. Ruvolo, N. Ekas, and A. Fogel. Applying Machine Learning to Infant Interaction: The Development is in the Details. *Neural Networks*, 23(8):1004–1016, 2010. 3

[22] S. Michelet, K. Karp, E. Delaherche, C. Achard, and M. Chetouani. Automatic Imitation Assessment in Interaction. In *Lecture Notes in Computer Science*, volume 7559, pages 161–173. Springer-Verlag, 2012. 3

[23] M. Pantic and A. Vinciarelli. Social Signal Processing. *The Oxford Handbook of Affective Computing*, page 84, 2014. 3

[24] A. Paxton and R. Dale. Frame-Differencing Methods for Measuring Bodily Synchrony in Conversation. *Behavior Research Methods*, 45(2):329–343, 2013. 3

[25] V. Petukhova and H. Bunt. Who's Next? Speaker Selection Mechanisms in Multiparty Dialogue. In J. Edlund, editor, *Workshop on the Semantics and Pragmatics of Dialogue*, pages 19–27. Royal Institute of Technology, 2009. 1

[26] F. Ramseyer and W. Tschacher. Nonverbal Synchrony in Psychotherapy: Coordinated Body Movement Reflects Relationship Quality and Outcome. *Journal of Consulting and Clinical Psychology*, 79(3):284–295, 2011. 2

[27] R. Schmidt, S. Morr, P. Fitzpatrick, and M. Richardson. Measuring the Dynamics of Interactional Synchrony. *Journal of Nonverbal Behavior*, 36(4):263–279, 2012. 3

[28] X. Sun, K. Truong, A. Nijholt, and M. Pantic. Automatic Visual Mimicry Expression Analysis in Interpersonal Interaction. In *Computer Vision and Pattern Recognition Workshops*, pages 40–46, 2011. 2

[29] X. Sun, K. P. Truong, M. Pantic, and A. Nijholt. Towards Visual and Vocal Mimicry Recognition in Human-Human Interactions. In *International Conference on Systems, Man, and Cybernetics*, pages 367–373, 2011. 3

[30] L. v. Swon. The Effects of Nonverbal Mirroring on Perceived Persuasiveness, Agreement with an Imitator, and Reciprocity in a Group Discussion. *Communication Research*, 30(4):461–480, 2003. 1

[31] J. Terven, J. Salas, and B. Raducanu. Robust Head Gestures Recognition for Assistive Technology. In *Lecture Notes in Computer Science*, volume 8495, pages 152–161. Springer, 2014. 6

[32] P. Wagner, Z. Malisz, and S. Kopp. Gesture and Speech in Interaction: An Overview. 57:209–232, 2014. 2

[33] P. Wittenburg, H. Brugman, A. Russel, A. Klassmann, and H. Sloetjes. Elan: A Professional Framework for Multimodality Research. In *Proceedings of Language Resources and Evaluation*, volume 2006, page 5th, 2006. 6

[34] X. Xiong and F. De la Torre. Supervised Descent Method and Its Applications to Face Alignment. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 532–539, 2013. 1, 3

[35] X. Yu, S. Zhang, Z. Yan, F. Yang, J. Huang, N. Dunbar, M. Jensen, J. Burgoon, and D. Metaxas. Is Interactional Dissynchrony a Clue to Deception? Insights from Automated Analysis of Nonverbal Visual Cues. *IEEE Transactions on Cybernetics*, 45(3):506–520, 2015. 3