Neurocomputing ∎ (∎∎∎∎) ∎∎∎–∎∎∎

Contents lists available at ScienceDirect

Neurocomputing

journal homepage: www.elsevier.com/locate/neucom

# Head-gestures mirroring detection in dyadic social interactions with computer vision-based wearable devices

Juan R. Terven [a,d], Bogdan Raducanu [b], María Elena Meza-de-Luna [c], Joaquín Salas [a,*]

[a] Instituto Politécnico Nacional, CICATA Querétaro, Cerro Blanco 141, Colinas del Cimatario, Querétaro, 76160, Mexico
[b] Computer Vision Center, Edifici "O" – Campus UAB, 08193 Bellaterra, Spain
[c] Universidad Autónoma de Querétaro, Cerro de las Campanas, Querétaro, 76010, México
[d] Instituto Tecnológico de Mazatlán, Corsario I 203, Mazatlán 82070, México

ARTICLE INFO

*Article history:*
Received 29 November 2014
Received in revised form
26 May 2015
Accepted 26 May 2015

*Keywords:*
Head gestures recognition
Mirroring detection
Dyadic social interaction analysis
Wearable devices

ABSTRACT

During face-to-face human interaction, nonverbal communication plays a fundamental role. A relevant aspect that takes part during social interactions is represented by mirroring, in which a person tends to mimic the non-verbal behavior (head and body gestures, vocal prosody, etc.) of the counterpart. In this paper, we introduce a computer vision-based system to detect mirroring in dyadic social interactions with the use of a wearable platform. In our context, mirroring is inferred as simultaneous head noddings displayed by the interlocutors. Our approach consists of the following steps: (1) facial features extraction; (2) facial features stabilization; (3) head nodding recognition; and (4) mirroring detection. Our system achieves a mirroring detection accuracy of 72% on a custom mirroring dataset.



## 1. Introduction

During face-to-face human interaction, nonverbal communication plays a fundamental role, as it is used to support the spoken message and to put special emphasis on certain aspects of it [1]. Usually, nonverbal communication is manifested through a multiplicity of behavioral cues including head movements, body postures/gestures, facial expressions, winks, tone of voice, verbal accent, and vocal utterances [2]. Sometimes, these cues are also known as social signals, a term coined by Pentland [3], because they are an undivided part of our social interaction. He was also the first one to claim that they could be quantified automatically to infer from them behavioral patterns in human interactions. The first attempt to prove this theory was reported in Curhan and Pentland [4], where the authors tried to predict the behavioral outcome of employment selection interviews using non-verbal audio cues. The same approach has also been applied for predicting salary negotiations [5] and speed-dating conversations [6]. Some research on behavior analysis during social interactions has focused on different aspects such as the role of participants in news broadcasts and movies [2,7], the detection of the leadership role during meetings [8,9], the inference of personality traits [10,11], and the simultaneous prediction of a job interview outcome and personality [12].

A very important aspect that takes part during social interaction is represented by mirroring, *i.e.*, when one interlocutor tries to mimic the attitude of the counterpart [13], by imitating speech patterns (accent, voice prosody), facial expressions, postures, gestures, and idiosyncratic movements. The study of mirroring has attracted the interest of psychologists for a long time [14]. Back then, the analysis was based on the manually annotation of videotapes for listener movements and the prosody of the accompanying speech. It has not been until recently that the study of mirroring captured the attention of the Human-Computer Interaction (HCI) community. A comprehensive recent survey can be found in Wagner et al. [15]. The *mirroring behavior* reveals very important information regarding participants' inter-personal states and attitudes and it represents a reliable indicator of cooperativeness and empathy during interaction. Recent research revealed that people who, even consciously, mimic the behavior of others activate behavioral strategies which may increase their chances to achieve their goals [16]. Thus, a social interaction presenting a high number of *mirroring behavior* is perceived to run more smoothly and the chances to reach a positive outcome or agreement increase significantly.

On the other hand, a backchannel is defined as a modality used by a listener to briefly intervene during the mainstream presentation given by a speaker in order to show his/her level of support with respect to the topic being discussed. The definition of a backchannel is relative and is strongly dependent on the

* Corresponding author.
*E-mail addresses:* j.r.terven@ieee.org (J.R. Terven), bogdan@cvc.uab.es (B. Raducanu), mezamariel@gmail.com (M.E. Meza-de-Luna), jsalasr@ipn.mx (J. Salas).

http://dx.doi.org/10.1016/j.neucom.2015.05.131
0925-2312/

particularity of the problem to be tackled. In the audio domain, backchannels are represented as linguistic vocalization such as *hmmm, aha, uhhh*, and *yeah*. In the visual domain, backchannels are usually associated with gestures such as smiling, winking, and head nodding; however, there could be more complex structures, such as facial expressions. Sevin et al. [17] used multimodal fusion of audio-visual backchannels in order to trigger the adequate response in an avatar. In a similar approach the objective was to compare the realization of the feedback signal in both an avatar and a physical embodied robot (AIBO) based on audio-visual backchannels fusion [18]. The improvement of head gestures detection (*nodding* in particular) has been done not only considering the visual information, but also taking into account the speaking status [19]. In addition, backchannels are usually studied in the context of *turn-taking* [20].

As the vast literature on the subject suggests, mirroring is a complex phenomena [13]. In this paper, we focus on the detection of head-nodding as a relatively simple non-verbal communication modality because of its significance as a gesture displayed during social interactions. According to social psychology, head nodding plays a very important role 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 intention for turn claiming [21,20]. Additionally, head nods can be used during speaker turns to elicit feedback from the listener [20]. The psychology literature suggests that the frequency of head nod events in face-to-face interactions can reveal personal characteristics or even predict outcomes. For instance, job applicants producing more head nods in employment interviews have been reported to be often perceived as more employable than applicants who do not [22,23]. In this sense, the ability to automatically detect head nods could be useful to build automatic inference methods of high-level social constructs. Therefore, in our context, mirroring will be inferred head noddings displayed by the interlocutors. As our results show even this reduced perspective of the problem is detect mirroring using a wearable device.

Our main contributions are:

1. We introduce an approach to extract a visual backchannel gesture (*i.e.*, nodding) from videos captured using wearable devices.
2. We develop a computer vision-based method to detect mirroring automatically, as an identical head gesture in a dyadic conversation, using a wearable device.
3. We provide a recorded custom database, representing a dyadic conversation setting, where each of the participants is wearing a pair of smart glasses. Ground-truth is provided by annotated head movements captured by a pair of fixed cameras facing each participant.

To our knowledge, this is the first time smart glasses equipped with a camera (see Fig. 2) have been used in such a setting.

The rest of the paper is structured as follows. Section 2 is dedicated to a review of the state-of-the art in the study of mirroring behavior both from a psychological and computational perspective. In Section 3, we describe the experimental setup and scenario definition used in our study. Section 4 presents our approach for automatic mirroring detection. In Section 5, we present the experimental results (both quantitatively and qualitatively, in terms of user experiences). Finally, we present our conclusion and provide guidelines for future work.

## 2. Related work

Perhaps because the study of mirroring can be approached by different scientific disciplines, we have found that several terms have been used to express its meaning. Pentland [24] stated that "mirroring occurs when one participant subconsciously copies another participant's prosody and gesture". This definition has been widely used to mean the display of similar postures while people interact with one another [25], and has been exemplified by situations where when "person *A* nods or smiles following person *B* who has nodded or smiled too" [26]. Interestingly, some researchers [27,26,28] give equivalent semantic meaning to mirroring and other words such as synchrony. However, in this research we use the definitions stated by Burgoon et al. [29] where mirroring involves visual behaviors identical in form, while interpreting synchrony as "a smoothly meshed coordination between the interactants". This section provides an overview of the *mirroring behavior* from the psychological and computational perspectives.

### 2.1. Psychological perspective

The study of mirroring in psychology, as a nonverbal behavioral process, dates back to the early 1970s. The researchers noticed that during a conversation, the parties involved exchanged both words and nonverbal cues, the latter as an effective form to adjust from each other in order to reach convergence in communication. Thus, early studies focused upon voice prosody such as accent imitation [30], vocal intensity [31], pause frequency [32], speaking rate [33], and speech patterns [34]. However, during a social interaction, people not only tend to imitate vocal features, but also to match each other's facial expressions and body gestures. For instance, LaFrance [35] found that listeners tend to mirror a speaker's posture whom they find engaging. Another research concluded that newborn babies [36] and adults [37] imitate facial gestures, while infants imitate vocalic sounds [38].

A comprehensive analysis of the role of mirroring in social interactions and how it affects our decisions and behavior can be found in Guéguen [16]. At a cognitive level, the *mirroring behavior* could be explained through the mind-body dualism, according to which, the mental processes are closely related to the body (the so-called relationship between thinking and action) [39]. In other words, the transition between mental states could be understood based on its analogy with the trajectory of a dynamic system through a series of space-states. From the point of view of evolution, mirroring happened long before human developed their linguistic capabilities, as a mechanism used by people to survive by helping them to communicate and coordinate better [40]. This explains why nowadays, due to the significant role played in modern society by social interaction, mirroring constitutes an automatic and unconscious act rooted in our brain [41]. Thus, seen from this social perspective, mirroring arose from the need to increase the social coherence among the members of a group and to feel a sense of psychological connection between themselves. In other words, individuals who were able to mimic each other had more opportunities to experience this psychological connection and would have had more probabilities to be kept within the community. The experimental evidence for this statement is ample. For instance, levels of mimicry are positively correlated with sales rates [42], helping behavior to explicit verbal solicitation [43], romantic interest [44], and success in patients undergoing psychotherapy [45]. Based on these research results, in our study, we establish a link between the detected mirroring and the level of satisfaction derived from the social interaction. This highlights the usefulness of the proposed method in sociological studies.

Please cite this article as: J.R. Terven, et al., Head-gestures mirroring detection in dyadic social interactions with computer vision-based wearable devices, Neurocomputing (2015), http://dx.doi.org/10.1016/j.neucom.2015.05.131

### 2.2. Computational perspective

Although psychological research on role analysis dates back to the early 1970s, the computational approach for studying mirroring behavior has only recently ed 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 could be found in Delaherche et al. [27] and Wagneret al. [15]. For the remaining of this section we review only the most relevant works related to our research.

Ramseyer and Tschacher [46] 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 global value of activity. Their experiments demonstrated that nonverbal synchrony is higher in genuine interactions contrasted with pseudo-interactions. A similar approach has been reported with a more sophisticated analysis [47]. 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.

The previous approach has been extended by incorporating also the nonverbal information contained in the audio channel [48]. More precisely, they extracted from the acoustic signal the following prosodic features: pitch, intensity, energy and speaking rate. Following a similar multimodal framework, Delaherche and Chetouani [49] studied synchrony on a cooperative task where both partners have to coordinate in order to build an assembled object. In order 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 for pitch and pause. On the other hand, regarding the visual domain, it appeared that the image of motion history is the feature that best captures the synchrony of actions. Also in a multimodal framework, Bilakhia et al. [50] 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.

Michelet et al. [51] presented an unsupervised method to estimate mirroring. For this purpose, they computed *Bag-of-Words* models [52] 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 [53,54]. Cross-spectral and relative phase analysis revealed that speakers' and listeners' movements contained rhythms that were not only correlated in time but also exhibited phase synchronization. Furthermore, the synchronization during these interactive sessions suggests that similar organizational processes constrain bodily activity in natural social interactions and, hence, have implications for the understanding of joint action in general. Barbosa et al. [55] used cross-correlation to measure the motion coordination of lips and tongue during dyadic conversations (based on audio-features). Ashenfelter et al. [56] studied the role of gender in the mimicry behavior. Their approach was based on a windowed cross-correlation measure to quantify the symmetry in head movement. The results revealed peaks of high correlation over narrow time intervals (2 s) and a high degree of nonstationarity, which was found to be related with the number of men in a conversation. Messinger et al. [57] 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 Facial Action Coding System (FACS) [58]. Two types of models have been used: a causal and a temporal one which were characterized in terms of turn-takings. A turn-taking was defined as a mother or infant transition that was immediately preceded by the transition of the other partner. In our case, gestures are recognized using 3D face models to detect orientation, followed by Hidden Markov Models to estimate behavior. Mirroring is measured with wearable cameras that each partner is using during the conversation.

## 3. Experimental setup and scenario description

In this paper, we address the problem of automatic mirroring detection in dyadic conversations using a computer-vision based approach. We have opted for head nodding as a backchannel to characterize the mirroring behavior. Our choice for computer vision is due to its non-invasive nature, which increases the chances of being accepted by people. In our opinion, this is a requirement in applications related with social interaction analysis, since it is based on people-centered technology. The main novelty of our approach is that we propose a solution based on wearable technology, *i.e.* smart glasses. These glasses possess a high-definition video-camera located in the bridge between the two lenses, as depicted in Fig. 2. The choice for this solution is motivated by the fact that such device offers a first-person perspective, compared to the classical fixed cameras, which offer a third-party perspective.

The experimental setup of our study consisted of two people engaged in a social interaction, sitting at a table in a face-to-face conversation, and wearing a pair of smart glasses. Additionally, we have installed a pair of fixed cameras on the table which face the participants. The role of these cameras is two-fold: (1) to compare the performance of the system against stable video and (2) to extract *ground-truth* for head nodding detection, since the video stream provided by the wearable cameras is contaminated with ego-motion.

We developed a conversational scenario in which a confederated psychologist interviewed students looking for academic orientation and support. The students were informed about the general aims and procedures of the research being carried out. In order to create a realistic conversation scenario, the students were instructed to ask a psychologist for advice regarding academic orientation and support. The conversations were conducted around three questions: (1) What to do in a given situation? (2) What is the psychologist's experience with similar problems? and, (3) How many sessions are needed? The confederated psychologist answered the questions in the same verbal style but controlling his nodding gestures in four occurrence levels, which constitute the experimental cases: *Low*, the psychologist acts trying not to make head gestures; *No-control*, the psychologist acts normally; *Confederated*, the psychologist acts mirroring the client's gestures; and *Promoted*, the psychologist acts trying to display more gestures than usual (see Fig. 7).

Please cite this article as: J.R. Terven, et al., Head-gestures mirroring detection in dyadic social interactions with computer vision-based wearable devices, Neurocomputing (2015), http://dx.doi.org/10.1016/j.neucom.2015.05.131

After the conversation, the students were asked to fill in a questionnaire in order to evaluate the usefulness (0=bad; 10=excellent) of the conversation in terms of: (1) the attention (understood as the sustained concentration of the psychologist on the problem); (2) whether they felt listened to (understood as the ability of the psychologist to pay attention to the student to hear what is being said); (3) the psychologist's competence (understood as the psychologist's capacity, skill, or ability to do the job correctly or efficiently); (4) the level of satisfaction in the interaction (understood as the perception of social interaction performance or outcome in relation to the expectations). In addition, the questionnaire included open-ended questions for comments and recommendations in order to improve the experience. Finally, the students had an interview to assess the satisfaction of the interaction.

## 4. Automatic mirroring detection

For automatic detection of the mirroring behavior, we rely on the results of our previous work. In Terven et al. [59], we introduced a robust head gesture recognition system based on multiple Hidden-Markov Models (HMMs). In the current work, we extend the experimental base of their use to measure their performance in the context of mirroring behavior (as a consecutive sequence of head noddings) for a wearable platform. In consequence, the procedure for automatic detection of mirroring consists of the following steps: (1) facial features extraction; (2) facial features stabilization; (3) nodding recognition; and (4) mirroring detection. For the remaining of the section, we will give a detailed explanation of each of these steps.

### 4.1. Facial features extraction

Our head gestures recognition is based on non-rigid face tracking with Active Appearance Models (AAM) [60], in which we estimate a set of facial features in each frame of a video stream. AAMs build a model that combines face appearance and shape from a set of aligned training data. For tracking purposes, this model is fit to the input image in order to find the position of the face. To fit the AAM to the input image we use the Supervised Descent Method (SDM) [61], which uses SIFT [62] to describe facial features (see Fig. 1).

### 4.2. Facial features stabilization

In order to recognize head gestures correctly from wearable videos, we need to apply a stabilization step, in order to compensate for 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 [63–65], in our case we only needed to stabilize the tracked facial features by compensating for camera motion. Fig. 3 illustrates a comparison of the facial features position from a static camera, a wearable camera, and the stabilized version. The three graphs display the facial features changes in the horizontal and vertical directions (orange and blue respectively). The first graph is from video taken with a static camera displaying the and end of head gestures (nods) as red vertical lines. The second graph shows the result of processing the frames in the same time interval taken with a wearable camera. In this graph, it is possible to appreciate that the facial features motion is highly contaminated by camera's ego-motion. These


**Fig. 2.** Smart glasses used as wearable technology. The glasses have embedded in them a high-definition camera in the bridge connecting the two lenses.

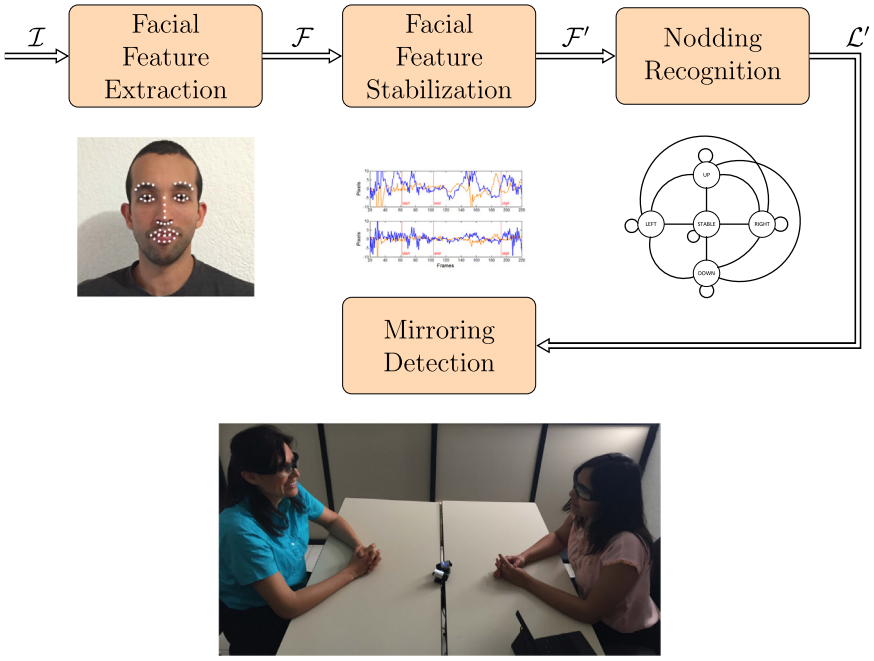


**Fig. 1.** Head-gestures mirroring detection using wearable devices in dyadic interactions. In our method, visual characteristics are extracted from an image stream $\mathcal{I}_k$ to both detect facial features $\mathcal{F}$ and compensate for egomotion $\mathcal{F}'$. The latter is necessary because the images are captured using wearable devices. Face activity $\mathcal{L}$ is modeled using a HMM, which in turn is analyzed temporarily to detect mirroring. The figures are described in detail in Sections 4.1–4.4 in the paper.

Please cite this article as: J.R. Terven, et al., Head-gestures mirroring detection in dyadic social interactions with computer vision-based wearable devices, Neurocomputing (2015), http://dx.doi.org/10.1016/j.neucom.2015.05.131

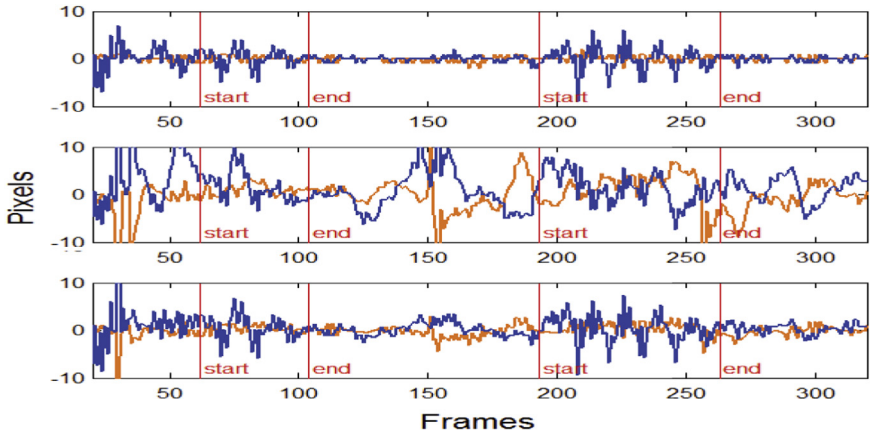


**Fig. 3.** Stabilized sequence (best seen in color). These graphs show the horizontal and vertical changes in position (in orange and blue respectively) of the head position in ten seconds of video. The first graph was taken from a static camera where vertical red lines indicate the start and end of a gesture. The second graph shows the same ten seconds of video from a wearable camera highly affected by camera motion. The third graph shows the stabilized sequence where the camera motion is attenuated while preserving head motion for gestures recognition. (For interpretation of the references to color in this figure caption, the reader is referred to the web version of this paper.)

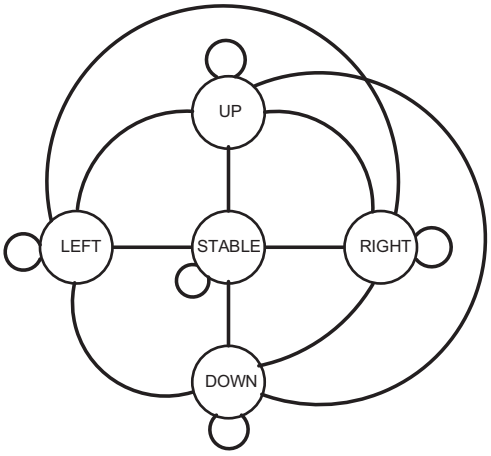


**Fig. 4.** Hidden Markov model (HMM) designed for head gestures recognition. The model states correspond to `LEFT`, `RIGHT`, `STABLE`, `UP`, and `DOWN` detected movements. Each gesture is recognized as the evaluation of a particular sequence of detected movements on a specifically trained HMM.

abrupt changes can trigger false detections of head gestures. The third graph shows the results for the same time interval frame from the wearable video but with motion stabilization. The stabilization process attenuates the camera motion while preserves the head motion for gesture recognition. Overall, the steps to stabilize the facial features are the following: (1) detecting and tracking of background features, (2) fitting a motion model for the camera, and (3) compensating for camera motion.

We estimate camera motion by detecting and matching SURF features [66] from the background in consecutive frames. We discard the SURF features from the face region to prevent the use of head motion as background motion. More formally, we create a set of $n$ matched keypoint pairs $(x_{t-1}^j, y_{t-1}^j)$ and $(x_t^j, y_t^j)$, for $j=1,\ldots,n$, from the previous and current frame. Using this set of keypoints, we estimate the interframe motion represented as a two-dimensional linear model with four parameters similar to the one proposed in Battiato et al. [63]:

$$\begin{aligned} x_t &= x_{t-1}\lambda\ \cos\ \theta - y_{t-1}\lambda\ \sin\ \theta + T_x, \\ y_t &= x_{t-1}\lambda\ \sin\ \theta + y_{t-1}\lambda\ \cos\ \theta + T_y, \end{aligned} \tag{1}$$

where $\theta$ is the rotation angle, $T_x$ and $T_y$ are the translation in the $x$ and $y$ directions, and $\lambda$ is a scale parameter. In order to estimate these parameters, we need at least four equations, so with two pairs of matched features it is possible to solve the system. However, due to noise in the coordinates of the features, it is more convenient to solve an over-constrained system of equations. Moreover, some of the features may have different motion due to wrong matches or moving objects in the background. To remove these outliers from the set of features, we use localized RANSAC as described by Grundmann et al. [64]. Once we remove the outliers, we are left with a set of $k$ feature pairs and solve for the four variables using linear Least Squares to get the motion matrix:

$$\mathbf{C} = \begin{bmatrix} \lambda\ \cos\ \theta & -\lambda\ \sin\ \theta & T_x \\ \lambda\ \sin\ \theta & \lambda\ \cos\ \theta & T_y \\ 0 & 0 & 1 \end{bmatrix}. \tag{2}$$

The non-stabilized position of the facial features $\boldsymbol{x}^w = [x^w, y^w, 1]^T$ can be described as the transformation due to the camera motion $\mathbf{C}$ of the static facial feature position $\boldsymbol{x}^s = [x^s, y^s, 1]^T$:

$$\boldsymbol{x}^w = \mathbf{C}\boldsymbol{x}^s. \tag{3}$$

In order to obtain a stable position of the facial features we apply the inverse transformation to the non-stabilize facial features:

$$\boldsymbol{x}^s = \mathbf{C}^{-1}\boldsymbol{x}^w. \tag{4}$$

### 4.3. Head gestures modeling

Following the approach described in Terven et al. [59], we created six fully connected Hidden Markov Models (HMMs) (see Fig. 4). Each HMMs is trained to recognize one of the following gestures: nodding, shaking, turning left, turning right, looking up, and looking down. Although we are interested only in nodding recognition, we still need a mechanism to discriminate between different gestures. During a conversation, many times, we display involuntary gestures which are not related with the context of the social interaction. For this reason, we need a robust nodding recognition system, which is not affected by false positives.

#### 4.3.1. Training

The goal of the training phase is to estimate the parameters of the HMMs (state transition and observation probability distributions) from data with known gestures. For this purpose, we used the head gestures dataset from Terven et al. [59] consisting in 30 annotated videos taken with a static camera and 10 annotated videos taken with a wearable camera. In total, this dataset contains 100 samples of each gesture. We used 70% of the gestures for training and 30% for validation. Each gesture in the training set is translated into a sequence of values containing the vertical and

Please cite this article as: J.R. Terven, et al., Head-gestures mirroring detection in dyadic social interactions with computer vision-based wearable devices, Neurocomputing (2015), http://dx.doi.org/10.1016/j.neucom.2015.05.131

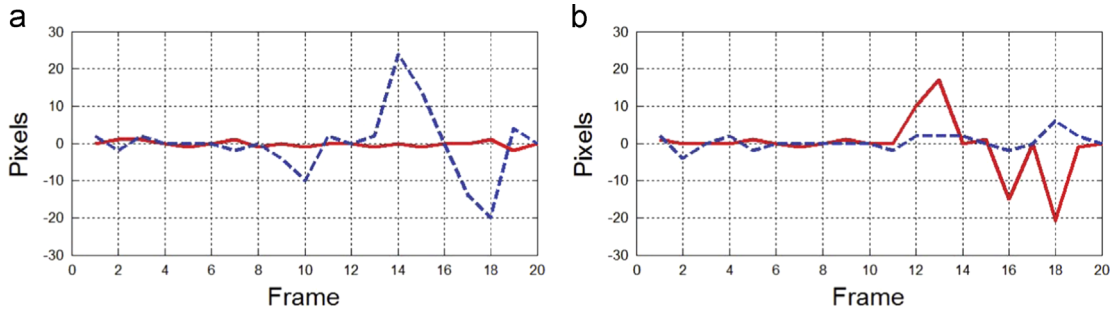


**Fig. 5.** Horizontal and vertical changes for typical nodding and shaking sequences. Solid lines represent the horizontal movement, dotted lines represent the vertical movement. (a) Motion changes in pixels for a nodding gesture. (b) Motion changes in pixels for a shaking gesture.

horizontal changes in consecutive frames. Fig. 5 shows a typical nodding and shaking sequences from our database. From these graphs, we see that a nodding gesture exhibits larger changes in the vertical direction than in the horizontal direction. Conversely, a shaking gesture exhibits larger changes in the horizontal direction than in the vertical direction. These clear distinctions make it easier to extract simple movements from the time series.

We defined five observation movements: `STABLE`, `UP`, `DOWN`, `LEFT`, and `RIGHT`. Each of these is coded with a number from 1 to 5. Then, for each training sequence, we extracted the observations from the horizontal and vertical changes using the following procedure: let $\Delta x$ represent horizontal change in two consecutive frames. Likewise, $\Delta y$ represents vertical change in two consecutive frames. If $|\Delta y| \gg |\Delta x|$ then the observation is `UP` or `DOWN` (*i.e.* 2 or 3) depending on the sign. Also if $|\Delta x| \gg |\Delta y|$ then the observation is `LEFT` or `RIGHT` (*i.e.* 4 or 5) depending on the sign. If none of these conditions are true, the observation is declared as `stable`. After this procedure, we are left with a sequence of values. For example, a nodding gesture is represented as a sequence of observations in the form $(1,2,2,1,3,3,1)$ after removing the `STABLE` repetitions. This example sequence stands for a `STABLE` observation followed by `UP` then `STABLE` followed by `DOWN`. To select the optimal threshold for discriminating between vertical and horizontal movements (the $\gg$ threshold), we tested different models in the validation set (each model with a different threshold). For each model, we calculated precision and recall and picked the model with the highest *F*-score.

#### 4.3.2. Recognition

Given an observation sequence extracted from video, the goal of recognition is to determine which one of the six HMMs is more likely to have generated the sequence. We used the Baum–Welch algorithm [67] to obtain the probabilities of the observation sequence given in each model. To determine the gesture, we selected the model with the highest probability. The head gesture recognition procedure is applied to the videos of both participants (both videos are synchronized in time) providing the following information: (1) type of gesture, (2) frame, and (3) end frame of each gesture.

### 4.4. Mirroring detection

Following an approach similar to Feese et al. [68], but measuring mirroring in both directions, we define 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 $\xi$ after person $\mathcal{B}$ started and within a time $\Delta t$ after person $\mathcal{B}$ stopped displaying gesture $\xi$. In case that person $\mathcal{A}$ displays $\xi$ multiple times while $\mathcal{B}$ is displaying $\xi$, only one event is counted. Similarly, a $m\mathcal{BA}$ event is triggered when person $\mathcal{B}$ starts displaying gesture $\xi$ after person $\mathcal{A}$ started and within $\Delta$ after person $\mathcal{A}$ stopped displaying gesture $\xi$. Gestures repetitions are 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 $t_1\left(g_i^{\xi}\right)$ and $t_2\left(g_i^{\xi}\right)$ respectively. An $m\mathcal{AB}$ event is triggered if (following Feese et al. [68]):

$$
\begin{aligned}
& g_i^{\mathcal{A}} = g_j^{\mathcal{B}}, \\
& \quad t_1\left(g_i^{\mathcal{A}}\right) > t_1\left(g_j^{\mathcal{B}}\right), \\
& \quad t_1\left(g_i^{\mathcal{A}}\right) < t_2\left(g_j^{\mathcal{B}}\right) + \Delta t.
\end{aligned} \tag{5}
$$

Fig. 6 shows a fragment of 33 s from one of the videos in our dataset. The top row depicts the nodding gestures performed by person $\mathcal{A}$. The middle row depicts the nodding gestures performed by person $\mathcal{B}$. Finally, the bottom row, depicts the mirroring behavior. The first three mirroring behaviors are triggered by the $\mathcal{A}$ person. As we could see 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 $\Delta t$. The other mirror events occur just after person $\mathcal{B}$ started the nodding gesture. The fourth mirroring behavior is due to the person's $\mathcal{B}$ response to the nodding gesture triggered by $\mathcal{B}$. The window $\Delta t$ is heuristically determined taken into consideration the analysis of our dataset where the average elapse time between gestures is 1.36 s.[1]

## 5. Experimental results

In this section, we first offer a detailed description of the collected dataset. After that, we report the performance of automatic mirroring detection and the performance in a real-world experiment. The experiment measures linear relationships between multiple scores of a social interaction and mirroring (ground truth and automatically detected).

### 5.1. Mirroring dataset

Based on the scenario presented in Section 3, we created a mirroring dataset consisting of 48 sessions of three minutes each, on average. The sessions were divided in four levels of mirroring (*Low, No-control, Confederated, Promoted*) of the confederated psychologist as described in Section 3. Fig. 7 shows the distribution of the controlled gestures in each case. After performing an Analysis of Variance test (ANOVA), we found significant differences between the *Low* and all the other cases with $p < 0.01$, and between *No-control* and *Promoted* cases with $p = 0.029$. However, we found no significant difference between the No-control and

[1] This estimated time could help as a reference value for a HCI system with the purpose to provide timely feedback to the user.

Please cite this article as: J.R. Terven, et al., Head-gestures mirroring detection in dyadic social interactions with computer vision-based wearable devices, Neurocomputing (2015), http://dx.doi.org/10.1016/j.neucom.2015.05.131

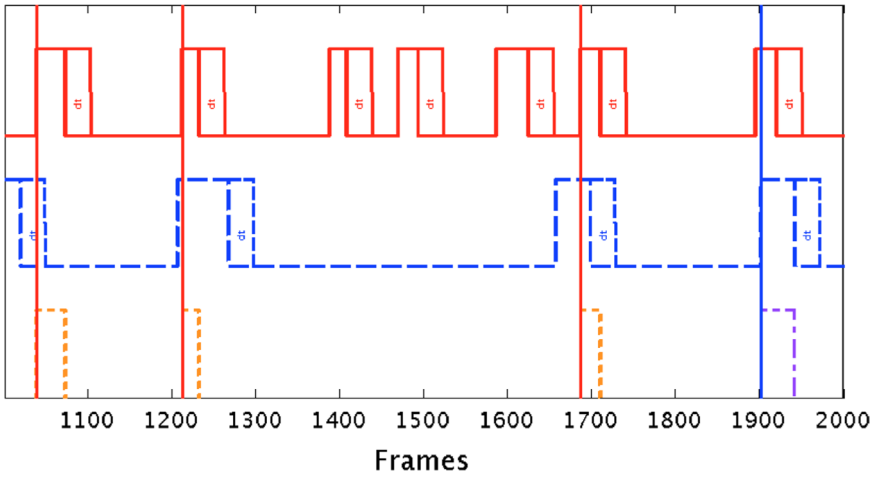


**Fig. 6.** Mirroring detection. Here we illustrate how the mirroring detection takes place. Rows one (red) and two (blue) show the number of frames when gestures from either person $\mathcal{A}$ or person $\mathcal{B}$ occur. Note that there is a fixed interval of time *dt* when the mirroring effect may take place. The third row displays the occurrence of mirroring. In frames 1050, 1210, and 1690 person $\mathcal{A}$ mirrored person $\mathcal{B}$ (orange). In frame 1900, person $\mathcal{B}$ mirrored person $\mathcal{A}$ (purple). (For interpretation of the references to color in this figure caption, the reader is referred to the web version of this paper.)

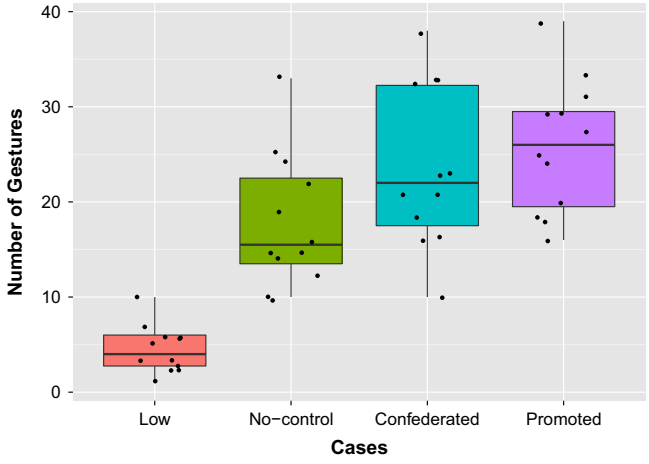


**Fig. 7.** Experimental cases displayed by the confederated psychologist. We show the distribution of gestures by case *Low*: the psychologist acted trying to make no head gestures. *No-control*: The psychologist acted naturally. *Confederated*: the psychologist acted mirroring the client's gestures. *Promoted*: the psychologist acted displaying more gestures than usual.

Confederated and also no significant difference between Confederated and Promoted. This lack of difference between adjacent cases may indicate a smooth transition between these classes.

To avoid biased results using the same person in different experimental conditions, we used a different person acting as student for each session. The sample size used for testing the algorithm described in Section 4 corresponds to the number of mirroring events in the whole experiment: 273 for the participants and 175 for the psychologist.

Our inclusion criteria consisted of the following requirements: (1) being at least 18 years old and (2) signing a participation agreement. The sample consisted of 48 volunteers (50% women), ranging from 18 to 44 years ($\mu$=21.00, $\sigma$=4,45), and college students (29.2% majoring in sociology, 25% in politics, 18.8% in architecture, 10.4% in engineering,6.3% in journalism, 4.2% in business, 4.2% in mathematics, and 2.1% in nursing).

We recorded each session with two static HD cameras and two wearable cameras. For the static cameras, we used Microsoft LifeCam Studio cameras fixed on the table looking at each participant. For wearable cameras, we used Pivothead glasses. After recording each session, we edited the four videos in order to synchronize the gestures in all the videos. Fig. 8 shows a snapshot of each synchronized video. Three trained sociology students annotated the starting and ending times of the nodding gestures in the videos using ELAN Linguistic Annotator [69]. These annotations served as gestures' ground truth and we used them to calculate the mirroring ground truth following the approach described in Section 4.4.

### 5.2. Automatic mirroring detection

The mirroring detection algorithm is based on head gestures recognition. For this reason, we address first the results related to this latter aspect. Fig. 9 shows the precision and recall curves of the head gestures and the mirroring detection algorithms for both, the static and wearable camera videos. We obtained these curves by changing the sensibility of the observation ($|\Delta y| \gg |\Delta x|$ and $|\Delta x| \gg |\Delta y|$) used for training the head gestures recognition described in Section 4.3.1. When sensibility varies, this has two effects: on one side, if the system is very sensible it will achieve high recall but low precision; on the other side, if the system is less sensible it will achieve higher precision but low recall. To find the optimal sensibility, we use the *F*-score defined as $F = 2 \times (precision \times recall)/(precision + recall)$.

Fig. 9(a) shows that the performance of the gestures recognition in the static-camera videos is higher than the performance in the wearable-camera videos. This is explained in terms of the ego-motion residuals, in spite of the video stabilization step. We have identified two main sources of miss detections or false negatives: (1) fast changes in head motion cause head tracking losses and (2) the stabilization algorithm smooth out subtle gestures. In Fig. 9 (b) we see that the performance of mirroring detection is affected by the performance of the head gestures recognition. Table 1 shows the precision and recall values of mirroring detection for each of the four experimental cases. We can see that the performance is not only affected by the type of camera but also by the amount of head gestures occurring in each case.

#### 5.2.1. Real-world performance

Precision–Recall curves show us the performance of the system as a function of true detections (true positives), false detections (false positives), and miss-detections (false negatives). However, how does this performance affect a real-world situation? For this, we analyze the data gathered from questionnaires that each participants answered after the session, and performed multiple

**Fig. 8.** Session videos. Here we illustrate the field of view while recording the videos. The top row shows an example frame from the wearable cameras. The bottom row shows an example frame from the static cameras. A video showing a single session is available at http://youtu.be/Ru7QSQVSu5s.

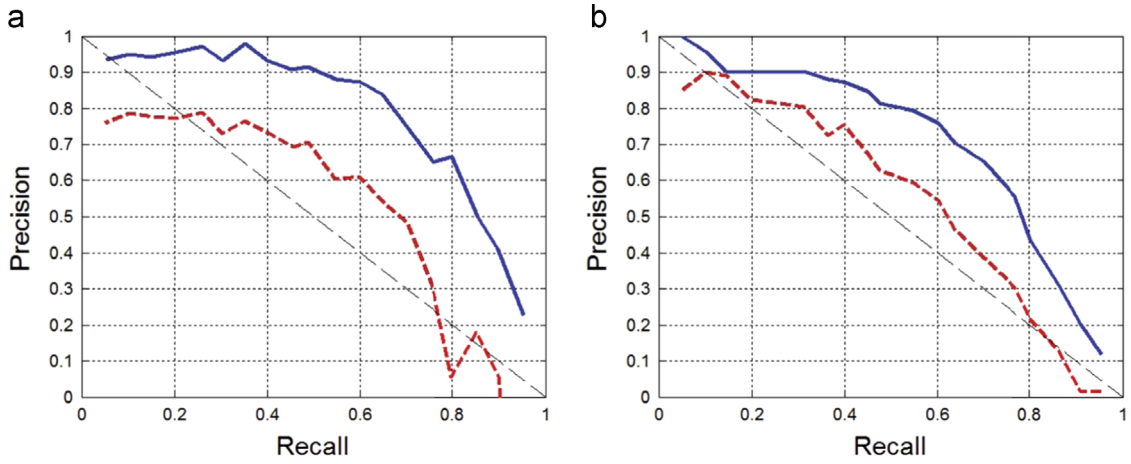


**Fig. 9.** Performance curves. Solid lines represent the performance using the static cameras, dotted lines represent the performance using the wearable cameras. (a) Precision and recall of head gestures recognition. (b) Precision and recall of mirroring detection.

**Table 1**
Case specific precision and recall measures of mirroring detection in the static-camera videos and wearable-camera videos.

| Case | Static camera | | | Wearable camera | | |
|---|---|---|---|---|---|---|
| | Precision | Recall | *F*-score | Precision | Recall | *F*-score |
| *Low* | 65.4 | 76.4 | 70.47 | 52.4 | 55.5 | 53.90 |
| *No-control* | 67.1 | 77.8 | 72.05 | 54.6 | 57.9 | 56.20 |
| *Confederated* | 68.5 | 78.2 | 73.02 | 55.3 | 59.6 | 57.37 |
| *Promoted* | 68.9 | 79.3 | 73.73 | 55.8 | 60.2 | 57.91 |

**Table 2**
Pearson Correlation *r* results between ground truth head nods and mirroring and the scoring of the interaction.

| *Scores* | $\mathcal{A}$ Nods | | $\mathcal{B}$ Nods | | *m* $\mathcal{AB}$ | | *m* $\mathcal{BA}$ | |
|---|---|---|---|---|---|---|---|---|
| | *r* | *p* | *r* | *p* | *r* | *p* | *r* | *p* |
| *Attention* | −0.08 | 0.57 | 0.09 | 0.55 | 0.04 | 0.79 | 0.19 | 0.20 |
| *Listening* | −0.16 | 0.27 | 0.11 | 0.45 | 0.00 | 0.99 | −0.03 | 0.82 |
| *Satisfaction* | 0.11 | 0.45 | 0.29* | 0.04 | 0.26 | 0.08 | 0.34* | 0.02 |
| *Competence* | 0.08 | 0.60 | 0.28 | 0.05 | 0.23 | 0.11 | 0.40** | 0.00 |

* Correlation is significant at the 0.05 level (2-tailed).
** Correlation is significant at the 0.01 level (2-tailed).

correlation tests to determine linear relationships between the amount of gestures or mirroring from ground truth and the scoring of the interaction. Then, we use the automatic detections in both, the static and wearable cameras instead of the ground truth to determine if the correlations hold. Table 2 shows the correlations between ground truth gestures, mirroring, and the scoring in the interaction. $\mathcal{A}$ Nods refer to participant's nodding, $\mathcal{B}$ Nods refer to psychologist's nodding, *m* $\mathcal{AB}$, is the client's mirroring, and *m* $\mathcal{BA}$ is the psychologist's mirroring. *r* is the Pearson correlation and *p* is the statistical significance. For these tests, we use a significance level of $\alpha = 0.05$.

Table 2 shows three statistical significant relationships marked with ∗: (1) there is significant small positive relationship between the amount of $\mathcal{B}$ nods (*i.e.* psychologist's nods) and the level of satisfaction in the interaction, $r(46) = 0.29, p = 0.04$; (2) there is a stronger positive relationship between the amount of *m*$\mathcal{BA}$ mirroring (*i.e.* psychologist mirroring participants) and both, the level of satisfaction in the interaction $r(46) = 0.34$, $p = 0.02$, and (3) the competence of the psychologist perceived by the participants $r(46) = 0.40$, $p = 0.004$. However, the received *attention* and

Please cite this article as: J.R. Terven, et al., Head-gestures mirroring detection in dyadic social interactions with computer vision-based wearable devices, Neurocomputing (2015), http://dx.doi.org/10.1016/j.neucom.2015.05.131

**Table 3**
Pearson Correlation $r$ results between the automatically detected mirroring from the static and wearable cameras and the scoring of the interaction.

| *Scores* | Static cameras | | | | Wearable cameras | | | |
|---|---|---|---|---|---|---|---|---|
| | $m\mathcal{AB}$ | | $m\mathcal{BA}$ | | $m\mathcal{AB}$ | | $m\mathcal{BA}$ | |
| | $r$ | $p$ | $r$ | $p$ | $r$ | $p$ | $r$ | $p$ |
| *Attention* | 0.02 | 0.88 | 0.19 | 0.18 | 0.04 | 0.82 | 0.12 | 0.50 |
| *Listening* | −0.00 | 0.95 | −0.01 | 0.91 | −0.11 | 0.52 | −0.25 | 0.16 |
| *Satisfaction* | 0.25 | 0.09 | 0.34* | 0.02 | 0.25 | 0.16 | 0.27 | 0.12 |
| *Competence* | 0.23 | 0.11 | 0.39** | 0.00 | 0.27 | 0.12 | 0.40* | 0.02 |

* Correlation is significant at the 0.05 level (2-tailed).
** Correlation is significant at the 0.01 level (2-tailed).

*listening* do not seem to be correlated neither with head nodding nor mirroring.

We repeat these tests using the automatically detected gestures and mirroring events from the static-camera videos and the wearable-camera videos. The small positive correlation between the amount of $\mathcal{B}$ nods and the level of satisfaction in the interaction was lost due to false negatives. However, the two significant correlations that relate mirroring with satisfaction and competence hold when using automatically detected mirroring from the static cameras and the correlation between the mirroring and competence also hold when using automatically detected mirroring from the wearable cameras. Table 3 shows these results.

## 6. Discussion

Egocentric vision, *i.e.* computer vision embedded in wearable devices, made possible by the miniaturization of video cameras, opened a new dimension in computer vision applications. Thanks to it, we are able now to address the same classic problems from a fresh perspective. *First-person* perception is very different from the previous *third-party*, represented by cameras located in our environment. For instance, objects and events do not appear isolated in the scene, but they could be analyzed in the user's context. For example, a particular object of interest for the user will be well-positioned in the camera's field of view, thus making the automatic image processing algorithms more robust and easier to overcome challenges represented by poor illumination, cluttered background, low image resolution, and so on.

Our choice for *smart glasses* is motivated by the need to have a user-centered perceiving device that resembles the point-of-view of a normal sighted person, which is a significant requirement for social interaction. Other existing wearable solutions (*e.g.*, SenseCam [70], or more recent Narrative Clip [71]) present the disadvantage that they should be hung around the neck, such that the camera is at chest level. This presents the impediment that they look always forward and capture images which are not related with what the person is looking at in a given moment.

From the perspective of the use of smart glasses for automatic social interaction analysis, several attempts have been performed. Krishna et al. [72] created a wearable system for face recognition which is robust to different face orientations and changes in illumination conditions. Gade et al. [73] proposed a robust person localization system based on the same technological platform. Fathi et al. [74] proposed a wearable system for the long-term analysis of social interactions. During one-day experiment, they tried to identify the relative head poses of nearby persons in different settings: street, amusement park and social events. The location and orientation of faces are estimated and used to compute the line of sight for each face. The context provided by all the faces in a frame is used to convert the lines of sight into locations in space to which individuals attend.

We believe that our solution could extend the use of computer-vision based wearable devices to the field of assistive technology. Having a system that could provide an automatic analysis of non-verbal communication during social interaction, would be of great benefit for people with visual impairment or suffering from the Autism Spectrum Disorder (ASD). Although, currently, the solution we propose is for mirroring detection only, it has the potential to be adapted and enhanced with new functionalities in order to serve as an assistive technology for the categories of people mentioned before. Perhaps, our approach could complement other wearable devices developed to address this problem [75].

## 7. Conclusion and future work

In this paper, we presented a computer vision-based approach for automatic detection of mirroring in dyadic social interactions using wearable devices. We have inferred the mirroring from visual backchannels represented by head-noddings. The method has been validated on a custom mirroring dataset. We have presented a thoroughly quantitative evaluation of users' experiences for the method described. Our experiments showed a significant correlation between the amount of mirroring and participants' satisfaction during the social interaction.

Regarding future work, we have identified three directions. First, we will look for an improved video stabilization algorithm in order to increase the recognition performance of head noddings. Second, we will extend the set of behaviors that our system is able to mimic in terms of head gestures and facial expressions. Third, we will be looking to test our approach in a real-world application. Motivated by the positive results obtained from our qualitative analysis, we will target most likely the domain of assistive technology.

## Ethics during the study

This study followed ethical standards as stipulated by the American Psychological Association [76]. An informed consent process was held. Confidentiality and person's anonymity were maintained at all times. All video and audio recordings were done with participant's written authorization.

## Acknowledgements

This work was partially supported by FOMIX GDF-CONACYT under Grant no.189005, by IPN-SIP under Grant no. 20150281, by UCMexus, and by MINECO Grants TIN2013-41751 and TIN2013-49982-EXP, Spain. Juan Ramón Terven was partially supported by Tecnológico Nacional de México and CONACYT. The authors are grateful with Roy Rajan for his comments to the document. Joaquin Salas is on sabbatical leave at FI-UAQ supported by CONACYT under Grant 234093.

## References

[1] M. Knapp, J. Hall, Nonverbal Communication in Human Interaction, Cengage Learning, Boston, USA, 2009.
[2] A. Vinciarelli, M. Pantic, H. Bourlard, Social signal processing: survey of an emerging domain, Image Vis. comput. 27 (12) (2009) 1743–1759.
[3] A. Pentland, Social signal processing, IEEE Signal Process. Mag. 24 (4) (2007) 108–111.

[4] J. Curhan, A. Pentland, Thin slices of negotiation: predicting outcomes from conversational dynamics within the first 5 min, J. Appl. Psychol. 92 (3) (2007) 802–811.
[5] R. Caneel, Social signaling in decision making, Master Thesis, MIT, 2005.
[6] A. Madan, R. Caneel, A. Pentland, Voices of attraction, in: Proceedings of the Augmented Cognition, 2005, see TR584, ⟨http://hd.media.mit.edu⟩.
[7] C. Weng, W. Chu, J. Wu, Movie analysis based on roles' social network, in: International Conference on Multimedia and Expo, 2007, pp. 1403–1406.
[8] B. Raducanu, D. Gatica-Perez, Inferring competitive role patterns in reality TV show through nonverbal analysis, Multimed. Tools Appl. 56 (1) (2012) 207–226.
[9] D. Sanchez-Cortes, O. Aran, M. Schmid-Mast, D. Gatica-Perez, Identifying emergent leadership in small groups using nonverbal communicative cues, in: International Conference on Multimodal Interfaces, 2010, Article 39.
[10] J. Staiano, B. Lepri, N. Aharony, F. Pianesi, N. Sebe, A. Pentland, Friends don't lie – inferring personality traits from social network structure, in: Ubicomp, 2012, pp. 321–330.
[11] B. Lepri, S. Ramanathan, K. Kalimeri, J. Staiano, F. Pianesi, N. Sebe, Connecting meeting behaviour with extraversion: a systematic study, IEEE Trans. Affect. Comput. 3 (4) (2012) 443–455.
[12] L. Nguyen, A. Marcos-Ramiro, M. Marron-Romera, D. Gatica-Perez, Multimodal analysis of body communication cues in employment interviews, in: International Conference on Multimodal Interfaces, 2013, pp. 437–444.
[13] T. Chartrand, J. Bargh, The Chameleon effect: the perception-behavior link and social interaction, J. Personal. Soc. Psychol. 76 (6) (1999) 893–910.
[14] W. Condon, W. Ogston, Speech and body motion synchrony of the speaker-hearer, in: The Perception of Language, Charles E. Merrill, 1971, pp. 150–184.
[15] P. Wagner, Z. Malisz, S. Kopp, Gesture and Speech in Interaction: An Overview, vol. 57, 2014, pp. 209–232.
[16] N. Guéguen, C. Jacob, A. Martin, Mimicry in social interaction: its effect on human judgement and behavior, Eur. J. Soc. Sci. 8 (2) (2009) 253–259.
[17] E. de Sevin, E. Bevacqua, S. Pammi, C. Pelachaud, M. Schröder, B. Schuller, A multimodal listener behavior driven by audio input, in: International Workshop on Interacting with ECAs as Virtual Characters, 2010.
[18] S. Al Moubayed, M. Baklouti, M. Chetouani, T. Dutoit, A. Mahdhaoui, J. Martin, S. Ondas, C. Pelachaud, J. Urbain, M. Yilmaz, Generating robot/agent backchannels during a storytelling experiment, in: IEEE International Conference on Robotics and Automation, 2009, pp. 3749–3754.
[19] L. Nguyen, J.-M. Odobez, D. Gatica, Using self-context for multimodal detection of head nods in face-to-face interactions, in: International Conference on Multimodal Interfaces, 2012, pp. 289–292.
[20] J. Allwood, L. Cerrato, A study of gestural feedback expressions, in: Nordic Symposium on Multimodal Communication, 2003, pp. 7–22.
[21] U. Hadar, T. Steiner, F. Clifford, Head movement during listening turns in conversation, Nonverbal Behav. 9 (4) (1985) 214–228.
[22] R. Gifford, C. Ng, M. Wilkinson, Nonverbal cues in the employment interview: links between applicant qualities and interviewer judgments, Appl. Psychol. 70 (4) (1985) 729–736.
[23] T. McGovern, B. Jones, S. Morris, Comparison of professional versus student ratings of job interviewee behavior, J. Couns. Psychol. 26 (2) (1979) 176–179.
[24] A. Pentland, Socially aware, Comput. Commun. Comput. 38 (3) (2005) 33–40.
[25] S. Feese, B. Arnrich, G. Troster, B. Meyer, K. Jonas, Detecting posture mirroring in social interactions with wearable sensors, in: IEEE International Symposium on Wearable Computers, 2011, pp. 119–120.
[26] X. Sun, K. Truong, A. Nijholt, M. Pantic, Automatic visual mimicry expression analysis in interpersonal interaction, in: Computer Vision and Pattern Recognition Workshops, 2011, pp. 40–46.
[27] E. Delaherche, M. Chetouani, A. Mahdhaoui, C. Saint-Georges, S. Viaux, D. Cohen, Interpersonal synchrony: a survey of evaluation methods across disciplines, IEEE Trans. Affect. Comput. 3 (3) (2012) 349–365.
[28] K. Bousmalis, M. Mehu, M. Pantic, Towards the automatic detection of spontaneous agreement and disagreement based on nonverbal behaviour: a survey of related cues, databases, and tools, Image Vis. Comput. 31 (2) (2013) 203–221.
[29] J. Burgoon, L. Stern, L. Dillman, Interpersonal Adaptation: Dyadic Interaction Patterns, Cambridge University Press, USA, 1995.
[30] H. Giles, Accent mobility: a model and some data, Anthropol. Linguist. 15 (1973) 87–105.
[31] M. Natale, Convergence of mean vocal intensity in dyadic communications as a function of social desirability, J. Personal. Soc. Psychol. 32 (1975) 790–804.
[32] J. Capella, S. Planalp, Talk and silence sequences in informal conversations. III.: Interspeaker influence, Hum. Commun. Res. 7 (1981) 117–132.
[33] R. Street, Speech convergence and speech evaluation in fact-finding interview, Hum. Commun. Res. 11 (1984) 149–169.
[34] S. Goldinger, Echoes of echoes: an episodic theory of lexical access, Psychol. Rev. 105 (1998) 251–279.
[35] M. LaFrance, Posture mirroring and rapport, in: M. Davis (Ed.), Interaction Rythms: Periodicity in Communicative Behavior, Human Sciences Press, New York, USA, 1982, pp. 279–298.
[36] A. Meltzoff, M. Moore, Newborn infants imitate adult facial gestures, Child Dev. 54 (1983) 702–709.
[37] G. McHugo, J. Lanzetta, D. Sullivan, R. Masters, Emotional reactions to a political leader's expressive displays, Journal of Personality and Social Psychology 49 (1985) 1513–1529.
[38] P. Kuhl, A. Meltzoff, Infant vocalizations in response to speech: vocal imitation and developmental change, J. Acoust. Soc. Am. 100 (1996) 2425–2438.
[39] D. Richardson, R. Dale, K. Shockley, Synchrony and swing in conversation: coordination, temporal dynamics and communication, in: Embodied Communication, 2008, pp. 75–93.
[40] J. Lakin, V. Jefferis, C. Cheng, T. Chartrand, The Chameleon effect as social glue: evidence for the evolutionary significance of nonconscious mimicry, J. Nonverbal Behav. 27 (2003) 145–162.
[41] J. Lakin, T. Chartrand, Using nonconscious behavioral mimicry to create affiliation and rapport, Psychol. Sci. 14 (4) (2003) 334–339.
[42] C. Jacob, N. Guéguen, A. Martin, G. Boulbry, Retail salespeople's mimicry of customers: effects on consumer behavior, J. Retail. Consum. Serv. 18 (5) (2011) 381–388.
[43] N. Guéguen, A. Martin, S. Meineri, Mimicry and helping behavior: an evaluation of mimicry on explicit helping request, J. Soc. Psychol. 151 (1) (2011) 1–4.
[44] S. Farley, Nonverbal reactions to an attractive stranger: the role of mimicry in communicating preferred social distance, J. Nonverbal Behav. 38 (2) (2014) 195–208.
[45] F. Ramseyer, W. Tschacher, Nonverbal synchrony of head- and body-movement in psychotherapy: different signals have different associations with outcome, Front. Psychol. 5 (979) (2014) 1–9.
[46] F. Ramseyer, W. Tschacher, Nonverbal synchrony in psychotherapy: coordinated body movement reflects relationship quality and outcome, J. Consult. Clin. Psychol. 79 (3) (2011) 284–295.
[47] X. Sun, K. Truong, A. Nijholt, M. Pantic, Automatic visual mimicry expression analysis in interpersonal interaction, in: Computer Vision and Pattern Recognition Workshops, 2011, pp. 40–46.
[48] X. Sun, K.P. Truong, M. Pantic, A. Nijholt, Towards visual and vocal mimicry recognition in human–human interactions, in: International Conference on Systems, Man, and Cybernetics, 2011, pp. 367–373.
[49] E. Delaherche, M. Chetouani, Multimodal coordination: exploring relevant features and measures, in: International Workshop on Social Signal Processing, 2010, pp. 47–52.
[50] S. Bilakhia, S. Petridis, M. Pantic, Audiovisual detection of behavioural mimicry, in: International Conference on Affective Computing and Intelligent Interaction, 2013, pp. 123–128.
[51] S. Michelet, K. Karp, E. Delaherche, C. Achard, M. Chetouani, Automatic imitation assessment in interaction, Lecture Notes in Computer Science, 7559, Springer, Berlin, Heidelberg (2012), p. 161–173.
[52] L. Fei-fei, P. Perona, A. Bayesian, Hierarchical model for learning natural scene categories, Comput. Vis. Pattern Recognit. 2 (2005) 524–531.
[53] R. Schmidt, S. Morr, P. Fitzpatrick, M. Richardson, Measuring the dynamics of interactional synchrony, J. Nonverbal Behav. 36 (4) (2012) 263–279.
[54] A. Paxton, R. Dale, Frame-differencing methods for measuring bodily synchrony in conversation, Behav. Res. Methods 45 (2) (2013) 329–343.
[55] V. Barbosa, M. Oberg, R. D'echaine, E. Vatikiotis-Bateson, An instantaneous correlation algorithm for assessing intra and inter subject coordination during communicative behavior, in: Workshop on Modeling Human Communication Dynamics, 2010, pp. 38–41.
[56] K. Ashenfelter, S. Boker, J. Waddell, N. Vitanov, Spatiotemporal symmetry and multifractal structure of head movements during dyadic conversation, J. Exp. Psychol. 35 (4) (2009) 1072–1091.
[57] D. Messinger, P. Ruvolo, N. Ekas, A. Fogel, Applying machine learning to infant interaction: the development is in the details, Neural Netw. 23 (8) (2010) 1004–1016.
[58] P. Ekman, W. Friesen, The Facial Action Coding System: A Technique for The Measurement of Facial Movement, Consulting Psychologists Press, Palo Alto, USA, 1978.
[59] J. Terven, J. Salas, B. Raducanu, Robust head gestures recognition for assistive technology, Lecture Notes in Computer Science, 8495, Springer International Publishing (2014), p. 152–161.
[60] T.F. Cootes, G.J. Edwards, C.J. Taylor, Active appearance models, IEEE Trans. Pattern Anal. Mach. Intell. 23 (6) (2001) 681–685.
[61] X. Xiong, F. De la Torre, Supervised descent method and its applications to face alignment, in: IEEE Conference on Computer Vision and Pattern Recognition, 2013, pp. 532–539.
[62] D.G. Lowe, Distinctive image features from scale-invariant keypoints, Int. J. Comput. Vis. 60 (2) (2004) 91–110.
[63] S. Battiato, G. Gallo, G. Puglisi, S. Scellato, SIFT features tracking for video stabilization, in: International Conference on Image Analysis and Processing, 2007, pp. 825–830.
[64] M. Grundmann, V. Kwatra, I. Essa, Auto-directed video stabilization with robust L1 optimal camera paths, in: IEEE Conference on Computer Vision and Pattern Recognition, 2011, pp. 225–232.
[65] Y. Matsushita, E. Ofek, W. Ge, X. Tang, H.-Y. Shum, Full-frame video stabilization with motion inpainting, IEEE Trans. Pattern Anal. Mach. Intell. 28 (7) (2006) 1150–1163.
[66] H. Bay, A. Ess, T. Tuytelaars, L.Van. Gool, Speeded-up robust features (SURF), Comput. Vis. Image Underst. 110 (3) (2008) 346–359.
[67] L. Rabiner, A tutorial on hidden Markov models and selected applications in speech recognition, Proc. IEEE 77 (2) (1989) 257–286.
[68] S. Feese, B. Arnrich, G. Troster, B. Meyer, K. Jonas, Quantifying Behavioral Mimicry by Automatic Detection of Nonverbal Cues from Body Motion, in: IEEE International Conference on Social Computing, 2012, pp. 520–525.
[69] P. Wittenburg, H. Brugman, A. Russel, A. Klassmann, H. Sloetjes, Elan: a professional framework for multimodality research, in: Proceedings of Language Resources and Evaluation, vol. 2006, p. 5.

Please cite this article as: J.R. Terven, et al., Head-gestures mirroring detection in dyadic social interactions with computer vision-based wearable devices, Neurocomputing (2015), http://dx.doi.org/10.1016/j.neucom.2015.05.131

[70] S. Hodges, L. Williams, E. Berry, S. Izadi, J. Srinivasan, A. Butler, G. Smyth, N. Kapur, K. Wood, SenseCam: a retrospective memory aid, in: UbiComp 2006: Ubiquitous Computing, Springer, Berlin, Heidelberg, 2006, pp. 177–193.
[71] M. Cardona, Designing the Tangible experience of interactive memories (Ph.D. thesis), Rochester Institute of Technology, 2013.
[72] S. Krishna, G. Little, J. Black, S. Panchanathan, Wearable face recognition system for individuals with visual impairments, in: Conference on Computers and Accessibility, Baltimore (MD), USA, 2005, pp. 106–113.
[73] L. Gade, S. Krishna, S. Panchanathan, Person localization in a wearable camera platform towards assistive technology for social interactions, in: Workshop on Media Studies and Implementations that Help Improving Access to Disabled Users, 2009, pp. 53–62.
[74] A. Fathi, J. Hodgins, J. Rehg, Social interactions: a first-person perspective, Comput. Vis. Pattern Recognit. (2012) 1226–1233.
[75] S. Boucenna, A. Narzisi, E. Tilmont, F. Muratori, G. Pioggia, D. Cohen, M. Chetouani, Interactive technologies for autistic children: a review, Cognit. Comput. 6 (4) (2014) 722–740.
[76] American Psychological Association, Publication Manual of the American Psychological Association, American Psychological Association, 2010.

**Juan R. Terven** is a doctoral student at Instituto Politécnico Nacional, Mexico. He works as a half-time lecturer at Mazatlan Institute of Technology (ITM). He has been a graduate visiting student at MIT and a research intern in Microsoft. Terven's research interests include embedded systems, computer vision, and assistive technologies design. He is a member of IEEE.

**Bogdan Raducanu** is a senior researcher and project director at the Computer Vision Center in Barcelona, Spain. His research interests include computer vision, pattern recognition, and social computing. He received a Ph.D. in computer science from the University of the Basque Country, Bilbao, Spain.

**María Elena Meza-de-Luna** is researcher at the UAQ, México. She is interested in the social and cultural issues to prevent the expression of violence. Currently, she is the director of !Atrévete Ya!/¡Hollaback!-Querétaro, an organization to prevent street harassment (http://www.atrevete-ya.org) and president of IIPSIS, an NGO devoted to research and intervention in psychosocial matters.

**Joaquín Salas** is a professor in the area of Computer Vision at Instituto Politécnico Nacional. His research interests include the development of assistive technology for the visually impaired and visual interpretation of human activity. He received a Ph.D. in computer science from Monterrey Institute of Technology and Higher Studies (ITESM), México.

Please cite this article as: J.R. Terven, et al., Head-gestures mirroring detection in dyadic social interactions with computer vision-based wearable devices, Neurocomputing (2015), http://dx.doi.org/10.1016/j.neucom.2015.05.131