This paper was published at the Int. Conf. on Multimodal Interaction, Santa Monica, October 2012

Using Self-Context for Multimodal Detection of Head Nods in Face-to-Face Interactions

Laurent Nguyen
Idiap Research Institute
EPFL
Switzerland
lnguyen@idiap.ch

Jean-Marc Odobez
Idiap Research Institute
EPFL
Switzerland
odobez@idiap.ch

Daniel Gatica-Perez
Idiap Research Institute
EPFL
Switzerland
gatica@idiap.ch

ABSTRACT

Head nods occur in virtually every face-to-face discussion. As part of the backchannel domain, they are not only used to express a 'yes', but also to display interest or enhance communicative attention. Detecting head nods in natural interactions is a challenging task as head nods can be subtle, both in amplitude and duration. In this study, we make use of findings in psychology establishing that the dynamics of head gestures are conditioned on the person's speaking status. We develop a multimodal method using audio-based self-context to detect head nods in natural settings. We demonstrate that our multimodal approach using the speaking status of the person under analysis significantly improved the detection rate over a visual-only approach.

Categories and Subject Descriptors

I.4.m [**Computing Methodologies**]: Image Processing and Computer Vision—*Miscellaneous*

Keywords

Social computing, head nods, face-to-face interaction, self-context, multimodal processing

1. INTRODUCTION

In face-to-face interactions, head nods occur in every discussion. In most cases, people producing the head nods are not even aware of the social signal they emit: head nods are often the result of automatic processes. Independently of their function or meaning, head nods can be defined as vertical up-and-down movements of the head, rhythmically raised and lowered [5].

The social psychology community was the first to examine the functions of head nods during face-to-face interactions. Apart from the obvious function of signaling a 'yes', head nods are used *inter alia* to display interest, enhance communicative attention by occurring in synchrony with the other's speech, or anticipate an attempt to capture the floor (*i.e.*, signaling a turn claim) [5, 1]. Head nods form a major mode of communication in *backchannelling*, that is, during listener turns [5, 1]. Additionally, head nods can be used during speaker turns to elicit feedback from the listener [1]. 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 [3, 7]. In this sense, the ability to automatically detect head nods could be useful to build automatic inference methods of high-level social constructs.

Most methods for automatically detecting head nods have been developed in the context of human-computer interaction (HCI). The primary goal of these studies was to enable a machine to detect a 'yes' signaled by a head nod. Within this context, some studies proposed to track interest points of the face [6, 13] and use state-based approaches such as finite state machines [2] and hidden Markov models [10, 13] to detect nodding. These approaches show good performance in restricted contexts where the head motions are explicit. However, these methods do not allow to detect subtle head movements that occur quite often in natural human face-to-face interaction.

Nodding does not occur in a void. It is known that the speaking status of people influences the dynamics of the displayed head gestures [5]. When a person is speaking, the motion of his head has typically greater amplitude, larger frequency range, and follows a close to random pattern [5]. On the other hand, when the person is listening, his head tends to be more static as a result of both attention to the speaker and the fact of being silent. In this sense, head gestures are multimodal: the dynamics are conditioned on the speaking status of the actor.

The contextual nature of nodding has been used in the recent past. Work addressing the prediction of backchannel feedback makes use of these findings. The goal of this line of research is to enable robots or conversational agents to produce natural backchannels. In this type of setting, the contextual information such as lexical information or prosodic cues are used to predict head nods [9].

Furthermore, communicative contextual audio-based features have been used to improve the detection of head nods in dyadic scenarios [10, 8]. In [10], the scenario consists of a human interacting with an embodied conversational agent; hence, head nods produced by the participant are not entirely natural. In [8], automatically and manually extracted

ICMI'12, October 22–26, 2012, Santa Monica, California, USA.
Copyright 2012 ACM 978-1-4503-1467-1/12/10 ...$15.00.

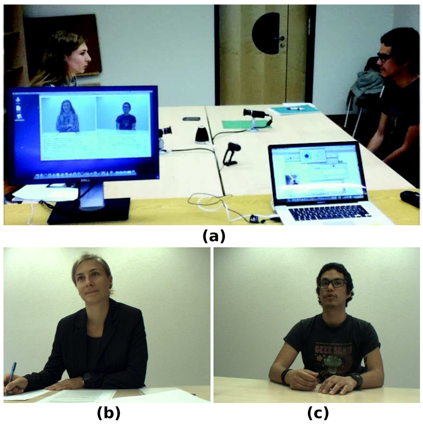

Figure 1: (a) Snapshot of the recording room. (b-c) Frame extracted from the video data for two protagonists in a dyad.

contextual cues (prosodic and lexical features) related to the speaker were used to improve the detection of the listener's head nods. In this constrained dyadic scenario, only one person spoke while the other person was asked to listen silently.

While the value of using audio-based context from the perspective of the speaker to improve the detection of listener head nods has been established, one aspect that to our knowledge has not been studied in detail is the effect of the audio-based self-context on head nod detection. This study develops a multimodal method using the self-context to detect head nods in fully natural conversations where both protagonists freely interact. We create an annotated dataset of natural dyadic interactions in order to benchmark our method. This paper is structured as follows. In Section 2, we present the dataset on which our method was trained and tested. In Section 3, we explain our method to automatically detect natural head nods. We discuss our results in Section 4 and conclude in Section 5.

2. NATURAL HEAD NOD DATASET

In order to benchmark head nod detection methods, we collected a dataset of 8 natural interactions (16 videos treated individually). Pairs of participants were asked to sit at both sides of a table and have a relaxed conversation on a topic of their choice. Dyads were acquainted before taking part to the experiment. In total, there were 9 different people (one person was in all conversations).

Video was recorded during the natural interactions. Illustrations of the experimental setup and the recorded data are displayed in Figure 1. Two 1280x960 monocular cameras were used, recording both protagonists of the dyad at 26.6 frames per second. Camera views were quasi-frontal, filming the upper part of the body. In addition to the images, the timestamps for each video frame were also recorded with a resolution of 1μs. In total, the dataset comprises approximately 160 minutes of recording (∼260'000 frames). Average video duration was 10 minutes.

In order to train and test the algorithm, annotations were performed on the dataset. Depending on the amplitude and duration of the up-and-down oscillatory movements, head nods can be difficult to code; two classes of head nods were therefore defined: *obvious* and *subtle*. Head nods were annotated by one of the authors, thus well acquainted with the concept of nodding, who noted the onset and offset time of an event, and qualitatively decided the nod class based on nod amplitude and duration. Speaking status was also manually annotated, marking the beginning and the end of a speaking segment. In total, nodding occurred during 858 seconds (22'812 frames), of which 92% occurred when the person under analysis was silent. Average head nod duration was 1.2 seconds. On average, speaking and silent times were split equally for each speaker in the dataset.

3. MULTIMODAL HEAD NOD DETECTION

As stated in the introduction, head nods are defined as vertical up-and-down movements of the head rhythmically raised and lowered. This implies an oscillatory pattern in the vertical axis, while the motion in the horizontal axis is limited. In order to encode this effect, we constructed features based on fine-grain motion detection and transformation into the frequency domain. The extraction of the features to characterize head nods follows a similar spirit than [10], but relies directly on the motion estimates derived from the video sequence rather than on the output state of a head tracker, which might not be so sensitive to subtle movements of the head. A binary classifier is then used to assign frames to one of two classes, *nodding* and *not nodding*.

3.1 Motion Estimation

Given the bounding box output of a face tracker (using the method described in [12]), the goal is to detect the motion in the horizontal and vertical directions. To perform this task, we used a parametric motion model which estimates the best set of parameters between the previous and current frames, using the face bounding box region.

We used the affine motion model defined in Equation 1, where (x_i, y_i) denotes a point in the image, $V(x_i, y_i)$ the flow vector modeled at point (x_i, y_i). Visual motion estimation over the whole face region provides an accurate estimation of head movements and therefore allows to capture subtle patterns using a multiresolution robust estimation method [11]. Parameters t_x, t_y, $a_{1:4}$ were estimated using least-mean-squares, implemented by the software package Motion2D[1].

$$V(x_i, y_i) = \begin{bmatrix} v_x(x_i, y_i) \\ v_y(x_i, y_i) \end{bmatrix} = \begin{bmatrix} t_x + a_1 x_i + a_2 y_i \\ t_y + a_3 x_i + a_4 y_i \end{bmatrix} \qquad (1)$$

We then computed the velocity at three arbitrarily defined points (see Figure 2) inside the bounding box, using the parameters of the optical flow model (Equation 1), providing the horizontal and vertical components of the motion at these three points. Roughly speaking, these points are around the mouth and eyes of the participant. Typical motion time-series for speaking, nodding while silent, and not nodding while silent are illustrated in Figure 2. The figure illustrates that head nod activity does present differences depending on the self-speaking status, and that building nodding models separately for the speaking and silent cases could reduce the confusion between, for instance, a head nod and a quasi-random head gesture displayed during a speaking turn.

3.2 Frequency Domain Analysis

Given the head motion in the horizontal and vertical directions, the goal is to capture the oscillatory characteristics

[1] http://www.irisa.fr/vista/Motion2D/

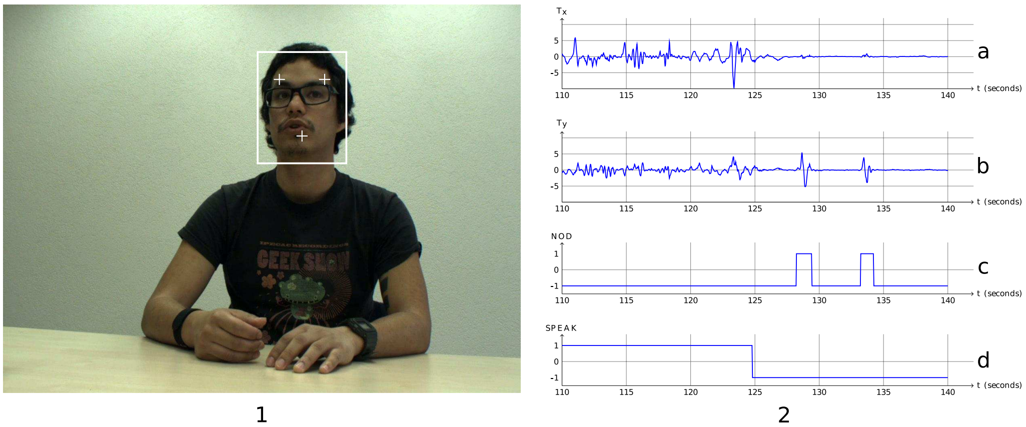

Figure 2: Illustration of the motion estimation step. In (1), the white rectangle is the face bounding box provided by the face tracker; the three white crosses are the pre-defined points where the motion is computed using the parametric model of Equation 1. A 30-second sequence of the motion is displayed in (2); (a) and (b) are the estimated motion in the horizontal and vertical directions, respectively; (c) shows the sequence of annotated nods ($1 =$ *nod*, $-1 =$ *non-nod*); (d) shows the speaking status of the participant ($1 =$ *speaking*, $-1 =$ *silent*).

of a head nod. In order to perform this task, we applied a Fourier transform (with Gaussian temporal window) to the velocity vectors $(v_x(x_i, y_i), v_y(x_i, y_i))^T$ of the pre-defined points, considering each vector component as an independent time-series. Although this independence assumption does not strictly hold, the computation is greatly simplified. Finally, the feature vectors are constructed by concatenating the Fourier transform outputs. Typical Fourier features for speaking, nodding while silent, and not nodding while silent are illustrated in Figure 3.

3.3 Classification

The goal is now to assign each feature vector to either of these two classes, *nodding* or *not nodding*. For each of the speaking status values, we trained a separate linear support vector machine (SVM) to perform the classification. This multimodal approach directly takes into account the switching dynamics of head movements, depending on the speaking status of the person under observation, as suggested by previous work in psychology [4].

The training set was defined as follows. The positive set was composed of all frames labeled as *obvious nods*. The negative set was selected randomly from the set of frames labeled as *non-nod*. Frames labeled as *subtle nods* were not used for training because they can be too similar to *not nodding* features. In addition to this, transitional frames were discarded to attenuate the dependence to time-related annotation inaccuracy. The training set was balanced, *i.e.* the number of positive and negative examples was equal. Approximately 5000 training examples were used for each class. We then further segmented the data into *speaking* and *silent*, training each separate model on its own training set.

To validate our hypothesis that self-context in terms of speaking status improves nodding detection, we implemented a baseline method using the visual modality only. For the visual-only method, we used a single SVM trained on the full training set (*i.e.* not separating it into *speaking* and *silent*).

4. RESULTS

The evaluation of the head nod detection method was conducted at the frame level. Leave-one-out cross validation was performed at the sequence level: the algorithm was trained on all except one sequence and tested on the remaining one. The binary output of the SVM classifier was compared to the annotated ground truth (including *obvious* and *subtle* nods).

In Figure 4, we display the receiver-operating characteristic (ROC) curve for both the visual-only and the multimodal approaches. The F_1 score, defined by the harmonic mean of precision and recall, was also computed. The multimodal method significantly outperforms the visual-only one: $F_1^{visual} = 0.559$, $F_1^{multi} = 0.6283$. These results show that using the self-speaking status of a person improves the detection rate, highlighting the difference in dynamics of head gestures between *speaking* and *silent* suggested in the psychology literature [5]. This result confirms previous findings [10, 8] that have shown the advantage of using audio-based contextual cues for this task, with the novel angle that using self-context (as opposed to interaction partner-based context) is also advantageous. Moreover, independently of the approach (multimodal or visual-only), the method we developed to extract head nods in natural settings yields competitive results on this dataset.

Head nods of small amplitude and duration are in general accurately detected by the proposed method. Additionally, because of the switching dynamics conditioned on the speaking status, the number of false positives is kept low. In other words, *speaking* can be seen as an attenuation factor of the head nod detector.

5. CONCLUSIONS

In this study, we developed and evaluated a multimodal method to detect natural head nods in face-to-face interactions. Our work brings the novel angle of examining in detail the effect of the speaking self-context on head nod detection. Two nodding models were trained, depending on the speaking status of the person under analysis. Compared to the baseline vision-only method, results demon-

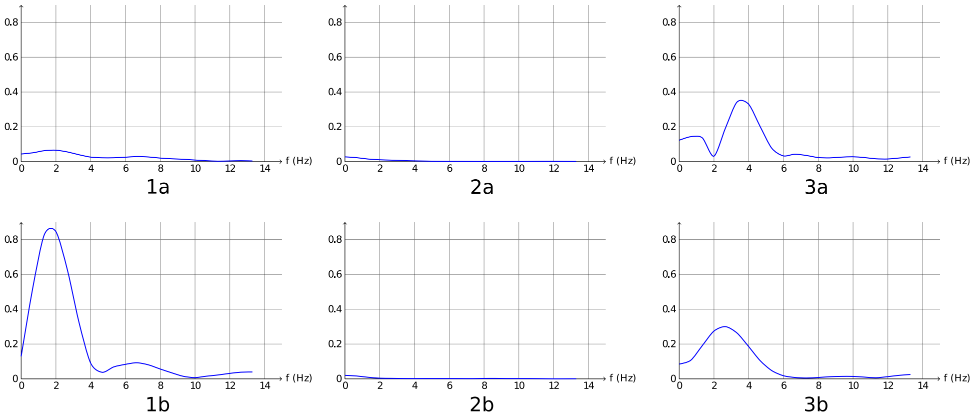

Figure 3: Examples of typical Fourier transform outputs. (a) and (b) are the Fourier outputs taken on the temporal motion sequence in the horizontal and vertical direction, respectively. Referring to Figure 2, (1) is a Fourier sample taken at $t = 128.8s$ (*nod*, *non-speaking*); (2) is sampled at $t = 131.4s$ (*non-nod*, *silent*); (3) is taken at $t = 116.2s$ (*non-nod*, *speaking*).

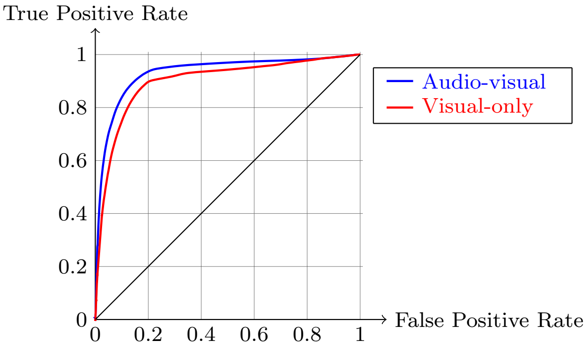

Figure 4: Receiver-operating characteristic curve (ROC) for the two head nod detection approaches. In red: visual-only method. In blue: audio-visual approach.

strated that audio-based self-context improved the detection of head nods, underlining the difference of head gesture dynamics conditioned on the speaking status of the person, as suggested by previous work in psychology. The developed method yielded competitive results on this dataset, allowing to detect subtle nods while keeping the number of false positive low.

The method presented in this paper could be extended for the detection of head shakes by using the same frequency domain motion features. Other possible avenues for future work would be to use more detailed audio context, *e.g.* taking into account the prosody or lexical information of both protagonists in the interaction. The presented method could also be seen as a first step towards the classification of head nods defined by their communicative functions; we hypothesize that context from both protagonists in the form of visual, prosodic, and lexical features would have to be used for this task.

Acknowledgments

This research was funded by SNSF SONVB project and the EC HUMAVIPS project. We thank Denise Frauendorfer (Université de Neuchâtel) for the support with data collection.

References

[1] J. Allwood and L. Cerrato. A study of gestural feedback expressions. In *Proc. First Nordic Symposium on Multimodal Communication*, 2003.

[2] L. Dong, Y. Jin, L. Tao, and G. Xu. Recognition of multi-pose head gestures in human conversations. In *Proc. Int. Conf. on Image and Graphics (ICIG)*, Aug. 2007.

[3] R. Gifford, C. F. Ng, and M. Wilkinson. Nonverbal cues in the employment interview: Links between applicant qualities and interviewer judgments. *Applied Psychology*, 70(4):729–736, 1985.

[4] U. Hadar, T. Steiner, E. Grant, and F. Rose. Kinematics of head movements accompanying speech during conversation. *Human Movement Science*, 2(1-2):35–46, June 1983.

[5] U. Hadar, T. J. Steiner, and F. Clifford Rose. Head movement during listening turns in conversation. *Nonverbal Behavior*, 9(4):214–228, 1985.

[6] A. Kapoor and R. W. Picard. A real-time head nod and shake detector. In *Proc. Workshop on Perceptive User Interfaces (ICMI-PUI)*, number 544, 2001.

[7] T. V. McGovern, B. W. Jones, and S. E. Morris. Comparison of professional versus student ratings of job interviewee behavior. *Counseling Psychology*, 26(2):176–179, 1979.

[8] L. Morency, I. de Kok, and J. Gratch. Context-based recognition during human interactions: automatic feature selection and encoding dictionary. In *Proceedings of the 10th international conference on Multimodal interfaces*, pages 181–188. ACM, 2008.

[9] L. Morency and I. D. Kok. Predicting Listener Backchannels: A Probabilistic Multimodal Approach. *Journal of Autonomous Agents and Multi-Agent Systems*, pages 1–14, 2008.

[10] L.-P. Morency, C. Sidner, C. Lee, and T. Darrell. Contextual recognition of head gestures. In *Proc. Int. Conf. on Multimodal Interfaces (ICMI)*, 2005.

[11] J. Odobez and P. Bouthemy. Robust multiresolution estimation of parametric motion models. *Visual Communication and Image Representation*, 6(4):348–365, 1995.

[12] E. Ricci and J. Odobez. Learning large margin likelihoods for realtime head pose tracking. In *Proc. Int. Conf. on Image Processing (ICIP)*, Nov. 2009.

[13] W. Tan and G. Rong. A real-time head nod and shake detector using HMMs. *Expert Systems with Applications*, 25(3):461–466, Oct. 2003.