Head Gesture Recognition in Spontaneous Human Conversations: A Benchmark

Yang Wu, Kai Akiyama
Nara Institute of Science and Technology, Japan
yangwu@rsc.naist.jp, akiyama.kai.ae6@is.naist.jp

Kris Kitani, Laszlo Jeni
Carnegie Mellon University, United States
kkitani@cs.cmu.edu, laszlo.jeni@ieee.org

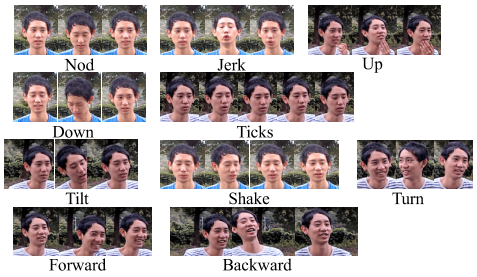
Yasuhiro Mukaigawa
Nara Institute of Science and Technology, Japan
mukaigawa@is.naist.jp

(a) Indoor (b) Outdoor (c) Wearable headset

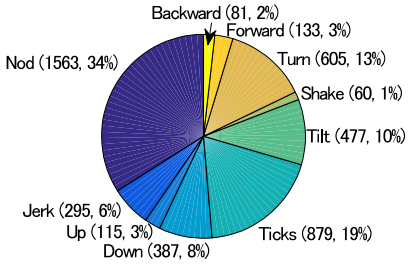
(d) Static camera videos (e) Wearable world camera videos (f) Eye camera videos (g) Gaze distribution

Axis	Gesture	Description
Pitch	Nod	Pitch down and up
	Jerk	Pitch up and down
	Up	Keep pitch up for a while
	Down	Keep pitch down for a while
	Ticks	Repetitive nods
Roll	Tilt	Tilt the head on one side
Yaw	Shake	Shake the head horizontally
	Turn	Rotate the head horizontally
Z (Scale)	Forward	Lean forward
	Backward	Lean backward

(h) Definition of head gesture types



(i) Examples of recorded head gestures



(j) Distribution of recorded head gestures

Figure 1. Overview of FIPCO. The first row shows how it was recorded. The second row presents sampled data from the three cameras and an accumulated gaze position plots over a whole sequence. The third row introduces the types of annotated gestures, including their definitions, exemplar data, and overall sample distribution (number of instances, and percentage over the samples of all gesture types).

Head gestures are important for transmitting and understanding attitudes and emotions in human face-to-face conversations. Existing approaches for head gesture recognition were only tested on small datasets with very few types (like "nod" and "shake") of performed gestures. It is thus unclear how well they can perform on spontaneous gestures in normal conversations, and there is also a lack of public and well-annotated benchmark datasets for the evaluation.

Here we introduce a novel spontaneous human conversation corpus built by ourselves, which is multimodal (video, audio, and eye gaze), consistent (collected under the same settings), gesture rich (4595 instances from 10 types), and

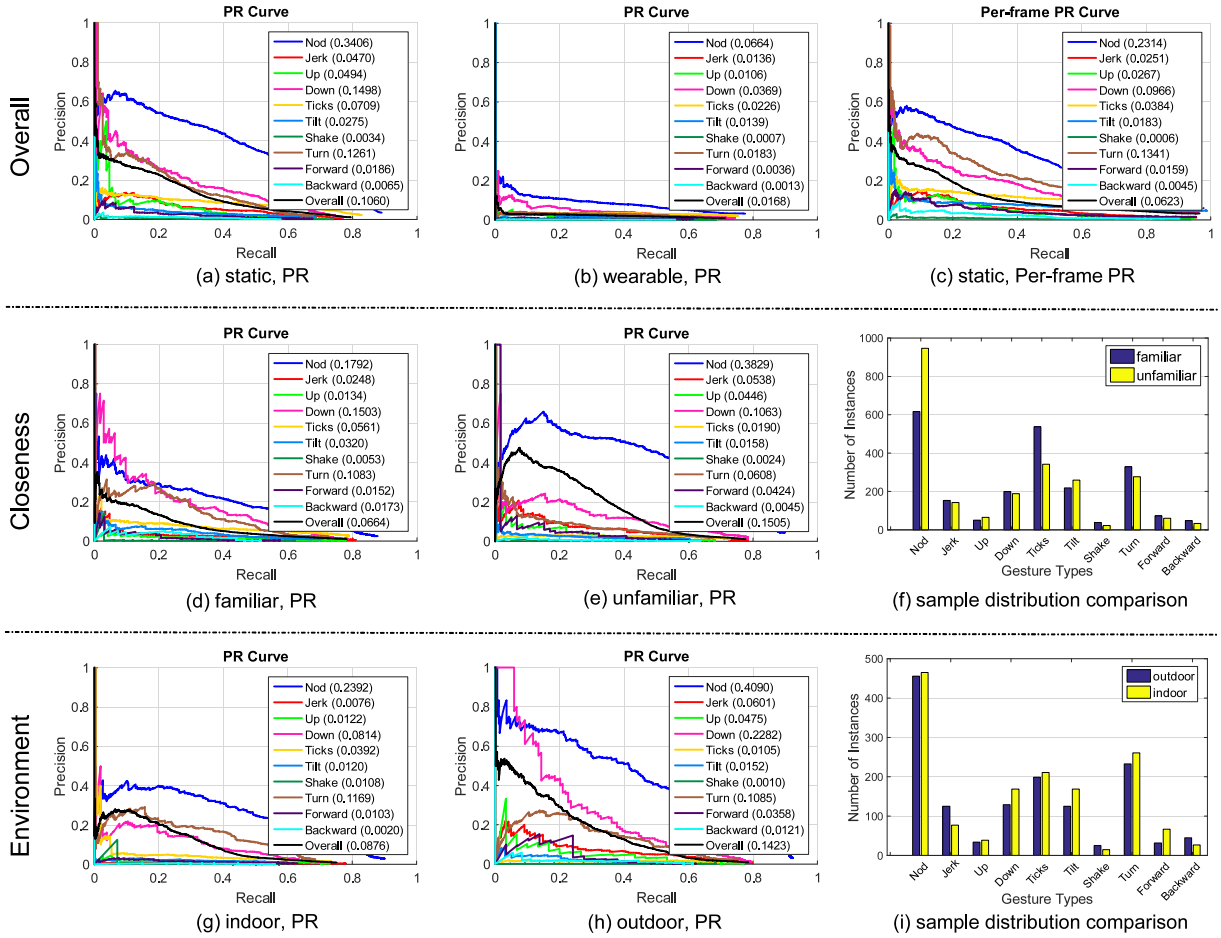


Figure 2. Experimental results of the baseline method: linear SVM based multi-scale sliding window search applied to the sequences of extracted head poses (output of Z-Face). All results were got from the leave-one-sequence-out splitting of training and test data. For a fair comparison of "indoor" and "outdoor" environmental settings, both of them use 10 sequences annotated by the same person. Results on "Closeness" and "Environment" are for the static camera data. The number following each gesture type is the average precision value.

multi-functional (with subsets for deeper studies on factors such as closeness of participants and scene of environment). More importantly, this corpus is **as far as we are aware the first one captured with first-person wearable cameras** as shown in Figure 1, though a static camera is also used for comparison. Therefore, it is named "First-Person Corpus (FIPCO)". We believe that first-person view is very important for understanding spontaneous human communication and interaction, and may enable many applications related to them. Some basic facts of FIPCO are listed in Table 1.

Table 1. Basic facts of FIPCO corpus

No. of participants	15
No. of pairs/conversations	30
No. of pairs per scene	10 indoor + 20 outdoor
No. of pairs per closeness	15 familiar + 15 unfamiliar
Length per conversation	$\sim$ 10 min.
Frame rate	24 fps
Total No. of frames	$\sim$1.3 million ($\sim$15 hours)

For automatic recognition, we provide head pose estimation results using the state-of-the-art 3D model based face tracker Z-Face, together with the eye tracking results of the Pupil Pro eye tracking headset. Some baseline results using only the video modal are given in Figure 2. Generally speaking, the baseline results (especially those for less frequent gesture types) are still far from satisfying. Recognizing (in fact detecting) spontaneous head gestures appears to be a very challenging task, and there is much room for improving, especially for the first-person view (which needs ego-motion removal). The corpus, together with a development toolkit and baseline source codes, will be made available to interested researchers.