Published in the 18th IEEE Int. Conference on Automatic Face and Gesture Recognition (FG'24), May 2024

CCDb-HG: Novel Annotations and Gaze-Aware Representations for Head Gesture Recognition

Pierre Vuillecard[1,2], Arya Farkhondeh[1,2], Michael Villamizar[1] and Jean-Marc Odobez [1,2]
[1] Idiap Research Institute, Martigny, Switzerland
[2] Department of Electrical Engineering, EPFL, Lausanne, Switzerland

***Abstract*— Despite remarkable progress in various human behavior perception tasks, head gesture recognition (HGR) has received limited attention in terms of datasets, benchmarks, and methods. In this work, we aim to address this gap and make two main contributions. First, we densely annotated the existing large-scale conversational dataset CCDb with diverse head gesture categories. This results in the CCDb-HG dataset, which can serve as a comprehensive benchmark for HGR research. Secondly, while previous gesture recognition methods have largely relied on head pose or facial landmarks as input, we propose to explore in addition the use of gaze to resolve ambiguous cases. This follows from the fact that head dynamics in interactions is driven by two main functions: communication (i.e. head gestures) and attention (i.e. gazing at other people or objects of interest). In fact, the head dynamics associated with attention activities can be confused for communication gestures, even though the gaze patterns are quite different in the two cases. In addition, we study several geometric and temporal data augmentation techniques to improve the generalization across novel viewpoints, as well as different model architectures to establish baseline performance on CCDb-HG. Our findings provide insights into various aspects of HGR and motivate further research in this field. To facilitate reproducibility, we will release the CCDb-HG annotations, code, and HGR models.**

I. Introduction

Nonverbal behaviors are a fundamental aspect of human communication, encompassing a wide range of non-linguistic cues that convey information beyond spoken words. These behaviors include facial expressions, body language, eye contact, vocal prosody, and other subtle cues that collectively play important roles in face-to-face settings [5]. Amongst them, head gestures play an important role in conversations where they express a multitude of essential functions related to speech production, turn talking, cognitive states, and emotions [24]. Indeed, head gestures convey meanings like agreement or disagreement, understanding or confusion, approval or disapproval, interest or boredom [9], [10]. Furthermore, head gestures can communicate nuanced information, adding depth and intricacy to interpersonal relationships [24].

However, despite their crucial role in communication, recognizing head gestures has not received adequate attention regarding comprehensive public datasets encompassing various real-life situations and contexts, standardized benchmarks and metrics for accurately evaluating the performance of HGR models, as well as novel models and methods. This paper makes a step in this direction, introducing significant dataset annotations, as well as a novel and more holistic approach for head gesture recognition.

The work was co-financed by Innosuisse, the Swiss innovation agency, through the NL-CH Eureka Innovation project ePartner4ALL (a personalized and blended care solution with virtual buddy for child health, number 57272.1 IP-ICT).

Fig. 1: Visual examples of the head gesture categories used in **CCDb-HG** (see Section III-A for their definition). Note that in communication gestures (nods, shakes, tilts), gaze is often fixated, which is not necessary the case of other gestures.

Datasets. Regarding datasets, several challenges persist. First, many existing datasets are not publicly available [2], [24], [26], [33]. Second, some datasets [2], [23] have only been limited to gesture labels like nods, which is not enough to represent the diversity of gestures found in real-world scenarios. Additionally, many datasets are small in scale [1], [2], [23], limiting their ability to capture the diversity of head gestures. To address these issues, we annotate all video recordings in an existing large-scale conversational dataset CCDb [1] with head gestures and introduce CCDb-HG. Compared to existing datasets, CCDb-HG is the largest publicly available head gesture dataset with diverse gesture

classes, as shown in Tab. I and Fig. 1. The release of CCDb-HG along with evaluation code will provide a valuable resource for the head gesture recognition community, enabling researchers to develop more advanced models and methods.

Methods. By definition, almost all existing methods for HGR primarily rely on sequences of head pose information (or facial landmark motion as proxy) as input [2], [11], [12], [18], [20]–[22], [25], [27], [29]. However, since head dynamics are naturally driven by gaze activities and communication, relying solely on head pose may lead to sub-optimal solutions, as certain head motion actions driven by attention can be misrecognized as communicative head gestures. For instance, a quick look downwards on a table and back to the interlocutor can be mistaken for a nod. Similarly, in a multiparty setting, a brief sideways glance from looking at a speaker to a second person (to check her reaction to the speech) and back could be incorrectly interpreted as a head shake. The source of confusion could be alleviated by exploiting gaze, since in communicative gestures (nods, shakes, tilt), gaze is often fixated, whereas in attention shifts, eyes and head motions are often coordinated (see Fig. 1). It is thus crucial to consider and model the interplay between head pose and gaze (eye) activities to avoid such confusions.

Generalization across expected data changes is another important factor to consider and is not so well studied as shown by the very limited amount of work that evaluates cross-dataset performance. For instance, models trained on purely frontal views (e.g., dyadic settings) may struggle with non-frontal views (e.g., group settings). To alleviate this issue, as head gestures are dynamic by nature, one may rely on motion cues and rate of changes [20]–[22], [24], [25], [27] rather than sequences of raw features (e.g. head pose angles), or even better, extract rotation invariant measures [2]. When using deep learning models, an alternative is to consider appropriate data augmentation techniques to increase data diversity and generalize to new settings.

In this paper, we explicitly study the benefits of data augmentation approaches. First, we introduce Geom-DA, a geometric augmentation technique explicitly designed to generate novel viewpoints for the input representation (pose, landmarks, gaze) of the head gesture recognition systems, hence promoting viewpoint invariance during model training and potentially removing the need for using specific features like dynamics. Secondly, we evaluate different temporal augmentation techniques, like standard perturbations used in time series analysis [30], but also consider a temporal version of the Mixup [7] data augmentation scheme, which has always improved results for image analysis tasks.

Finally, we investigate various deep learning models, encompassing 1D-CNN, GRU, LSTM, and temporal convolutional networks (TCN) [15], to benchmark them on CCDb-HG and explore their appropriateness for gesture recognition.

Contributions. Our results highlight that CCDb-HG offers a good benchmark for addressing the head gesture recognition task, presenting new opportunities for training deep learning models with good generalization capacity. In addition, we explore several factors (input, data augmentation, architecture) that can lead to effective head gesture recognition systems. In summary, our contributions are as follows:

- We densely annotated the CCDb dataset with several head gesture categories, leading to CCDb-HG, the largest head gesture dataset publicly available;
- We are the first to explore the influence of gaze as an auxiliary cue for head gesture recognition, demonstrating higher accuracy in general and more robustness;
- We investigate several factors for designing robust head gesture recognition methods (spatial and temporal data augmentation, need for invariant features, architecture) and throughly evaluate their impact on both within and cross-dataset settings.

Finally, by releasing our annotations, evaluation code, and HGR models, we aim to promote reproducibility, provide useful benchmarks, and stimulate further advancements in the field of non-verbal behavior analysis in the future.

II. Related Work

Head Gesture Datasets. Previous studies have primarily been evaluated within controlled laboratory settings [11], [12], [18], [29], where participants were directed to perform specific gestures, notably nods or shakes. To enrich this taxonomy, seminal works like [32] and [14] introduced coding schemes with five additional categories, contributing to defining a more comprehensive understanding of the diverse spectrum of head movements. Furthermore, endeavors to capture natural human interactions in ecologically relevant settings yielded datasets like NOMCO [26] and FIPCO [33]. Unfortunately, these datasets are not publicly available.

The Cardiff Conversation Database (CCDb) [1] stands out as a publicly accessible dataset with acceptable resolution and natural conversational content including head gestures. However, its existing annotations are limited both in size and gesture categories and, at times, quality. Leveraging CCDb, we have annotated the entire corpus with a diverse head gesture set, ensuring meticulous and high quality annotations through a refined protocol, resulting in CCDb-HG.

Head Gesture Methods. Typical framework consists of two distinct steps: feature extraction, and classification. Regarding the former, some works focused on tracking facial attributes like eye location and nose [11], [12], [18], [22], [25], [29] which is enough for capturing simple gestures like nods and shakes but faces challenges for other gestures like tilts and waggles and in generalizing to different viewing angles. With advancements in pose estimation [13], 3D head orientation became more extensively used [2], [20], [21], [24], [27], potentially combined with facial landmarks [27]. In our work, we enrich the representation of facial behavior by incorporating gaze as an additional modality, and unlike many previous works, we consider all inputs in 3D, including head pose, landmarks, and gaze, enabling a more generalizable representation across various dataset settings.

Additionally, many of the works have used relative differences such as velocity [2], [22], [24], [25]. Relative differences are particularly valuable for capturing motion patterns. However, in [27], they used position and velocity for

Datasets	#Videos	#F	#E	#S	#G	A
Ubimpressed [2]	11, 50min	10k	407	11	1	✗
FIPCO [33]	30, 15h	NA	NA	15	10	✗
NOMCO [26]	12, 12h	72k	3k	12	9	✗
KTH-Idiap [23]	9, 50min	3.6k	136	9	1	✓
CCDb [1]	30, 2h	16k	403	16	3	✓
CCDb-HG (ours)	115, 8h	178k	5k	22	6	✓

TABLE I: Head gesture datasets (#F: Frames, #E: Events, #S: Subjects, #G: Categories, A: Public availability).

the head pose and position only for the landmarks. We can argue that absolute position input contains richer information and that a model can learn to capture motion patterns. In our work, we investigate and compare the differences and potential advantages of both relative and absolute inputs. Building on the idea of invariant head pose features for robustness in diverse viewpoints introduced in [2], we extend this concept to landmarks and gaze, comparing it with a viewpoint data augmentation approach.

Data augmentation has received limited exploration in HGR. A notable exception is [27], which addressed data imbalance and limited diversity by generating additional gestures with varying speeds and scales, but did not present any ablation study. Here we explore the efficacy of similar standard time series augmentation [30] (TS-DA) and the temporal version of Mixup [7] (Mixup-DA).

Finally, various classifiers have been used for HGR. Early models included Hidden Markov Models [11], [12], [18], [29] and Support Vector Machine [22], demonstrating promising performance but struggling with scalability on large-scale datasets. Deep learning models, like Multi-Layer Perceptron [25], Convolutional Neural Network (CNN) [24], LSTM and Conv-LSTM architecture [27], have also been explored. In our work, we rely on these architectures to establish informative benchmarks on CCDb-HG, shedding light on their performance in challenging HGR scenarios.

III. The CCDb-HG head gesture dataset

In this section, we introduce the CCDb-HG dataset, which extends the full Cardiff Conversation Database (CCDb) dataset [1] with a dense and comprehensive set of gesture annotations, as described below.

A. The CCDb dataset

The CCDb dataset was collected to facilitate the detection and prediction of facial backchannel expressions and gestures. It comprises 49 non-scripted, natural dyadic conversations between pairs of individuals, recorded from a frontal viewpoint with a focus on the upper body. Only eight of these conversations have full annotations for speaker activity, facial expressions, head motion, and non-verbal utterances.

Annotations are available for three head gesture categories (Nod, Shake, and Tilt). However, it was observed that the quality of annotations was unsatisfactory with many gesture instances missing. Thus, to increase the size, diversity, and quality of annotated data in the CDDb dataset, we fully annotated 49 conversations with six head gesture categories.

B. Head Gesture Categories: Definitions

Defining head gestures is difficult, since, their form are sometimes not fully specific and can significantly vary amongst individual people, and because the same gestures can serve multiple functions [24]. As our aim was to go beyond nods and shakes while remaining at a satisfactory level of annotation agreement, we mainly follow the works in FIPCO [27], [33] and Kousidis et al. [14] with few modifications to define the categories. More specifically, we adopt the 7-coarse head gesture categories introduced in FIPCO, excluding categories involving body movements (forward/backward) and static head gestures (i.e. looking up or looking down). Instead, to focus on the recognition of dynamic head movements, we added the two following categories, namely, Waggle [14], [26], and Up/Down. The definition of the head gesture categories is as follows.

Nod is an up-down rotation along the pitch axis. It involves a slight, quick, or repetitive lowering and raising of the head. It comes under different variations in FIPCO [33], namely nod, jerk, and ticks. But as mentioned in [14], [27], those gestures are difficult to disambiguate for annotators.

Shake is a left-right horizontal rotation along the yaw axis. It involves a rapid and potentially repeated side-to-side motion, typically with small or moderate amplitude.

Tilt is a sideways rotation along the roll axis, involving a shift of the head in which one ear moves closer to the shoulder while the other ear moves away.

Turn corresponds to a left or right rotation, involving the shifting of the head from its original position to another one facing a different direction. Head turns can vary in amplitude, ranging from a slight turn to a complete reorientation of the head. It differentiates from a shake by being a nonrepetitive movement and often initiated by a gaze shift.

Waggle usually happens when speaking [14], [26], and involves a rhythmic swaying motion typically performed in a repeated manner. Unlike nod, shake, and tilt, waggle involves several head axis at the same time.

Up/Down is similar to a turn, but along the pitch direction and usually involves a gaze shift in the same direction as the head. Note that this definition differs from the Up/Down class in FIPCO, as it encompasses a dynamic movement and unifies up and down into a single category.

C. Annotation Protocol

Two annotators were employed to annotate the CCDb dataset using the head gesture definitions provided above. Each annotator was assigned 50 videos for annotations, as well as 15 videos which were annotated by both of them to evaluate the inter-annotator agreement. The annotation process comprised three stages: warm-up, dataset annotation, and review. Before the main annotation, a warm-up session was held where annotators practiced on four videos, received feedback and continued until they reached an understanding of the gesture classes. In the next phase, annotators annotated all videos, raising concerns via issue reports for specific segments. Finally, two domain experts reviewed the videos

Gesture	#F	#E	Frame F1	Event F1
Nod	97k	2469	0.80 (9k)	0.84 (250)
Shake	33k	848	0.83 (3.7k)	0.85 (91)
Tilt	14k	523	0.75 (1.7k)	0.71 (65)
Turn	16k	643	0.76 (2k)	0.81 (80)
Up/Down	6k	248	0.57 (0.6k)	0.63 (28)
Waggle	11k	192	0.06 (0.7k)	0.08 (12)
All	178k	4923	0.76 (17k)	0.79 (526)

TABLE II: Gesture distribution in CCDb-HG by number of frames (*#F*) and events (*#E*), with Frame F1 and Event F1 representing annotator agreement on 15 randomly selected videos (average annotated instances in parentheses). This results in a Cohen's Kappa coefficient of 0.75.

and annotations, addressing any annotator concerns to ensure annotation accuracy and reliability.

Besides category-specific definitions, general guidelines were given: gestures in the same category are segmented into different events only if there is a noticeable gap, otherwise, they are seen as one continuous sequence. In rare instances where gestures occur simultaneously, the first gesture, as ordered in Sec. III-B, is prioritized.

D. Analysis of Annotations

Tab. I underscores the unique attributes of CCDb-HG in contrast to other datasets. Besides its public availability, CCDb-HG distinguishes itself by offering a good compromise in terms of amount of subjects, gesture category, and data (duration, number of annotated events). Tab. II sheds light on the distribution of frames and events per class, revealing a known challenge namely, class imbalance: Nod (54.4% of frames and 50.0% of events) or Shake (18.5% of frames and 17.2% of events) occur more frequently than others, resulting in an uneven sample distribution.

As mentioned earlier, we assessed the inter-annotator agreement using 15 videos annotated by both annotators. We calculate F1 frame and event scores for each category. Results indicate a high level of agreement (overall Cohen's kappa score of 0.75), but the agreement for Up/Down is slightly lower, and that of Waggle is really low. From the inter-annotator confusion matrix (see Fig. A in sup. materials), most of the disagreement is due to gesture events being annotated by one person and not by the other, usually because they are rather subtle instances, and not so much due to inter-gesture confusion. The main exceptions is Waggle, which is often confused with Shake, Tilt, and None. Indeed, Waggle can sometimes be seen as a combination of these gestures, leading to discrepancies between annotators. Note that the number of instances of this category (12) in the 15 videos is rather low, so these results might not be significant. Nevertheless, in view of the low agreement, we decided to exclude the Waggle category in our experiments.

IV. Method

In this section, we first present an overview of our head gesture recognition system. We then detail the methods used for extracting the face related feature, as well as the different alternatives to build our input representation. Subsequently, we delve into the specificity of the deep network recognition models and finally introduce the data augmentation techniques aimed at enhancing generalizability and robustness.

Overview. Our recognition system is shown in Fig. 2. It takes as input a facial video clip v, comprising T consecutive frames f_t with dimensions $H \times W \times 3$. A feature representation I_t^e is created from the frame f_t by extracting multiple facial cues, including facial landmarks, head pose, and a gaze vector. Subsequently, the sequence of I_t^e is pre-processed to build the input representation I_t used as input to a trainable head gesture classifier C_θ.

A. Multiple Cue Extraction

To represent the face sequence, various head and face cues are extracted using the methods detailed below.

Head Pose. It is the primary cue for HGR [2], [27], providing information about the head rotational movements. To compute it, we extracted a set of 3D facial landmarks from the video frame t, using the Mediapipe landmark detector [19], and performed a Procrustes analysis between the extracted landmarks and a canonical set of facial landmarks (3D face model). This approach provides the orientation of the head pose $h_t \in \mathbb{R}^{1\times 3}$, that we encode as Euler angles (yaw, pitch, roll) expressed in camera coordinates.

Facial Landmarks. Head pose information can be compromised by the accuracy of its estimator, which can lead to noisy predictions in dynamic natural environments when people are talking or exhibiting strong facial expressions. To mitigate this, we exploit facial landmarks that can reinforce and complement head pose information. As specified above, we obtain 3D facial landmarks expressed in pixel units via Mediapipe [19] and keep only 5 of them (near ears, eyes, and nose) which are more immune to the above perturbation and can be estimated accurately over time [27]. Furthermore, as landmark positions in pixels are affected by resolution variations and subject-camera distance, we normalize the measures using the head size defined as the pixel distance between the 3D ear landmarks. We end up with a vector of 3D landmarks $l_t \in \mathbb{R}^{1\times 15}$.

Gaze. Some head movements are naturally driven by gaze activities. Consequently, we incorporate gaze information as an auxiliary cue to enhance the disambiguation of certain head gestures. We employ a ResNet50 network with a linear regressor, pre-trained on the ETH-XGaze dataset [34], to estimate 3D gaze direction. This network processes a normalized face patch as input, where normalization involves canceling head roll rotation and maintaining a fixed face-center to a virtual camera [34], [35], producing a normalized gaze direction g_t^n in spherical coordinates. This direction is then transformed back to the original camera coordinate system, producing $g_t = (\theta_t, \phi_t) \in \mathbb{R}^{1\times 2}$.

B. Input Representation

Altogether, our extracted features from a video clip v of T frames form the multivariate time-series $I^e \in \mathbb{R}^{T\times 20}$, encompassing head pose ($h \in \mathbb{R}^{T\times 3}$), facial landmarks

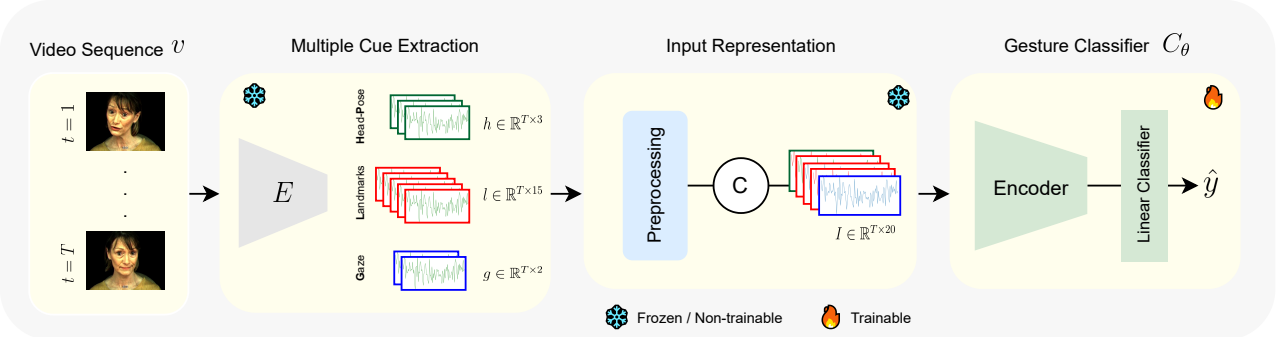


Fig. 2: Approach overview. A face video clip v consisting of T frames f_t is processed by different modules in E to extract multiple head cues (head pose, facial landmarks, gaze). Extracted features are then pre-processed to derive the clip representation I used as input of the gesture classifier C_θ. We explore different methods to handle variable viewpoints, like extracting invariance cues in the preprocessing step, or generating data for novel viewpoints (Geom-DA data augmentation).

($l \in \mathbb{R}^{T\times 15}$), and gaze ($g \in \mathbb{R}^{T\times 2}$). Given these cues, our goal is to pre-process these features to obtain an input representation according to $I = P(I^e)$ to achieve different goals and evaluate the impact of this processing on performance on different datasets. More specifically, the variants we investigate are named absolute (Abs), relative (Rel), and Invariance (Inv):

$$I^m = \{h^m, l^m, g^m\}, \quad \text{where, } m \in \{Abs, Rel, Inv\} \tag{1}$$

Abs. We simply have $I^{Abs} = I^e$. As I contains all information, in principle, deep models should be able to identify the patterns enabling the recognition of the different gestures under all conditions. However, this is true as far as these conditions are present in the data and we have enough data.

Rel. Relative differences are valuable for capturing dynamical patterns. By focusing on changes, they provide important information for identifying gestures while filtering out less relevant viewpoint dependent variations. Thus, it can be beneficial in low-data regimes, where I^{Abs} struggle to handle unseen views. Hence, to obtain I^{Rel}, we simply concatenate the relative values derived through simple channel-wise simple differences, i.e. $z_t^{Rel} = z_t - z_{t-\Delta}$ where $z \in \{h, l, g\}$, and Δ is a time difference that we set to 5 in practice

Inv. In real-world scenarios, head gestures can be observed from different viewing angles since people are not always facing the camera. Thus, a robust head gesture recognition system must be viewpoint invariant, which is not intrinsically the case when using I^{Abs} or I^{Rel} . Following [2] which addressed invariance for the head pose, our aim is to compute an input representation I^{Inv} that is invariant to the viewpoint. The main principle is to represent the different features (pose, landmarks, gaze) at time t w.r.t to the coordinate frame associated with the same features at time $t - \Delta$. For the head pose, we follow [2], in which the Euler angles of the head pose at time t in the coordinate frame of the head pose at time $t - \Delta$ is used as representation:

$$h_t^{Inv} = EA\left(R^T(h_{t-\Delta})R(h_t)\right) \tag{2}$$

where $R(h)$ denotes the rotation matrix associated with the Euler angles h, and EA is the inverse function that provides the Euler angles of a rotation matrix. For landmarks, we express their relative positions w.r.t. the head center p in a coordinate frame at time $t - \Delta$ by rotating these relative positions, and use their difference in position as representation:

$$\begin{aligned} l'_{t-\Delta} &= R^T_{t-\Delta}(l_{t-\Delta} - p_{t-\Delta}) + p_{t-\Delta} \\ l'_t &= R^T_{t-\Delta}(l_t - p_t) + p_t \\ l_t^{Inv} &= l'_t - l'_{t-\Delta} \end{aligned}$$

Finally, for gaze, we extent [2] as:

$$g_t^{Inv} = EA_{sc}\left(R^T(g_{t-\Delta})R(g_t)\right) \tag{3}$$

where rotations (and Euler angles) are defined from spherical coordinates in this case.

C. Data Augmentation

In this section, we introduce the different data augmentation schemes we have investigated to improve generalizability. These comprise Geom-DA, which can be considered as an alternative to address viewpoint robustness via geometric augmentation, and the temporal augmentation methods TS-DA and Mixup-DA to further boost robustness.

Geom-DA. In the pursuit of encouraging the network to be invariant to viewpoint, we introduce Geom-DA which aims at generating synthetic head gesture samples with novel orientation. More precisely, Geom-DA operates by applying the same rotation transformation on the entire sequence of 3D facial cues within a head gesture sample. This effectively changes the head orientation while keeping the same head motion. Importantly, to ensure realism and data balancing, we monitor the head pose distribution of generated samples so that it is close to the uniform distribution, and avoid random rotations which can lead to unrealistic poses. Then, given a realistic target head pose h_{target}, we modify a given input sample so that it has this target head pose on average. Accordingly, the augmented data sample is defined as:

$$\begin{aligned} h_t^{aug} &= EA\left(R(h_{target})R^T(\hat{h})R(h_t)\right) \\ l_t^{aug} &= R(h_{target})R^T(\hat{h})(l_t - p_t) + p_t \\ g_t^{aug} &= EA_{sc}\left(R(h_{target})R^T(\hat{h})g_t^u\right) \end{aligned}$$

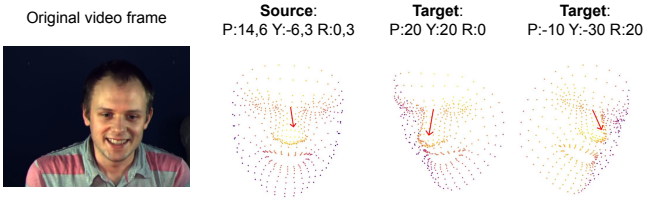


Fig. 3: Visual examples of Geom-DA.

where $\hat{h}$ denotes the mean of the head pose Euler angles of all T frames of the data sample, and g_t^u expresses g_t as a unit-vector. Fig. 3 depicts visual examples of Geom-DA.

TS-DA. To perform data augmentation in the temporal domain, we followed the typical approach of [30] by defining a set of transformations involving jittering, scaling, magnitude warping, and time warping. It should bring diversity like inter-person gesture variation or noise in the extraction. These transformations are applied sequentially with different probabilities (see Sec. 4 in sup. materials).

Mixup-DA. Mixup [7] is a popular data augmentation technique in the visual domain. The main idea behind Mixup is to linearly combine pairs of examples and their labels from the training data to create new synthetic examples. Mixup-DA randomly selects two training examples and linearly interpolate their input representation, alongside their labels to yield a new synthetic example. The intuition is that this can encourage the model to learn more generalizable features that are robust to small structured variations in the input.

D. Gesture Classification

Several deep network architectures have been investigated as classifier C_θ to process the input representation I and predict the gesture class $\hat{y}$. In all cases, we assumed the availability of a labeled dataset $\mathcal{D} = \{v_i, y_i\}$, and trained the classifier by minimizing the following loss:

$$\mathcal{L} = \frac{1}{N} \sum_{i=1}^{N} \ell(C_\theta(I_i), y_i) \tag{4}$$

where ℓ is the cross entropy loss. In all cases, our classifier comprised an encoder and a classifier head consisting of linear layer taking the embedding produced by the encoder and appling the Softmax function to output the gesture class. The following architectures were used as encoders.

CNN. 1D Convolutional Neural Networks is a prevalent choice for time-series classification, obtaining impressive results across numerous multivariate time-series benchmarks [8]. Our configuration closely follows the specifications outlined in [31], with the exception of using 128 channels in all three layers.

RNN. Recurent Neural Network has been an adequate choice for time-series modeling due to its capacity to leverage the dynamical characteristics of the data [28]. Here we evaluated both Gated Recurrent Unit (GRU) [3] and bidirectional Long Short-Term Memory (LSTM) [3], [6], [27] which enables modeling longer temporal dependencies. We rely on standard architectures with two layers and a hidden dimension of 64.

TCN. Temporal Convolutional Network (TCN) received significant attention in time-series modeling [4], [16], [17]. It is a 1D-CNN variant with a better ability to capture temporal patterns and model long-range dependencies with a minimal number of parameters (hence avoiding overfitting) thanks to the use of kernel dilation. Our implementation follows the Single-Stage TCN model proposed in [4]: we incorporate four layers to achieve a receptive field of $r = 2^{4+1} - 1 = 31$, which proves sufficient for our temporal input size.

V. Experiments

A. Evaluation Protocol

Datasets: We experiment using CCDb-HG and KTH-Idiap [23]. CCDb-HG features dyadic sessions with mainly frontal views, while the KTH-Idiap dataset features groups of four persons discussing around a table, so that looking at others involves more head motion. We divide the CCDb-HG dataset into train and test sets based on subject-level splitting, with four subjects (S_2, S_5, S_{20}, and S_{10}) assigned to the test set, comprising 25 videos out of the total 115 videos. KTH-Idiap comprises 9 videos (one per subject), necessitating leave-one-person cross-validation for evaluation. The original version of this dataset exclusively includes the Nod class. Consequently, we re-annotated KTH-Idiap dataset using the same procedure as in CCDb-HG, outlined in Section III-C.

Metrics and statistical tests: Performance is evaluated with frame-based and event-based F1 scores (see Sec. 3 in sup. materials). We report micro (overall) and macro (average class) measures. Due to the class imbalance, minority class improvements stand out more in macro. For the main table highlighting our contributions (Tab. VI and Tab. VII), we report the mean and standard deviation of 5 runs. To evaluate the significance of differences between methods, we perform a one-sided T-tests, with the null hypothesis assuming differences due to randomness. In section V-C, we report the significance level with * and ** for p-value below 0.05 and 0.01 respectively.

B. Implementation Details

Training Samples: We create samples from videos using a 31-frame window, with the central frame determining the label. To avoid noisy samples, we exclude 7 frames before and after a gesture's onset and offset, similar to [2]. Negative samples are taken from non-gesture windows with a 7-frame gap between them to reduce redundancy. This process yields 180,568 samples for CCDb-HG, 53% being non-gesture.

Input Normalization: Input normalization involves subtracting the per-sample channel-wise mean and dividing by the standard deviation computed from all samples.

Training Details: CNN models are trained for 30 epochs using the Adam optimizer with a 0.0005 learning rate, 0.0001 weight decay, and cosine decay scheduler, utilizing a batch size of 128. To address class imbalance, predominantly from the None class, we employ a focal loss with a gamma of 1.0. For selecting the best hyperparameters, a grid search was performed on the learning rate, hidden dimension, filter size, and loss. The best parameter over a 3-fold cross-validation on the training split has been selected.

Input Representation	Event-based F1							Frame-based F1						
	Nod	Shake	Tilt	Turn	Up-Down	All-Micro	All-Macro	Nod	Shake	Tilt	Turn	Up-Down	All-Micro	All-Macro
CCDb-HG														
Head-Pose	0.53	0.64	0.47	0.50	0.10	0.52	0.45	0.47	0.59	0.38	0.46	0.08	0.47	0.40
Landmarks	**0.78**	0.65	0.52	0.53	**0.23**	0.68	0.54	**0.74**	0.64	0.43	0.48	0.12	0.67	0.48
Head-Pose + Landmarks	0.77	**0.72**	**0.55**	0.51	0.20	0.68	**0.55**	**0.74**	**0.68**	**0.46**	0.48	**0.13**	**0.68**	0.50
Head-Pose + **Gaze**	0.54	0.61	0.44	0.59	0.15	0.53	0.47	0.49	0.57	0.37	0.57	0.12	0.49	0.42
Landmarks + **Gaze**	**0.78**	**0.72**	0.46	0.62	0.18	0.68	**0.55**	**0.74**	**0.68**	0.41	**0.59**	0.11	**0.68**	**0.51**
Landmarks + Head-Pose + **Gaze**	**0.78**	0.70	0.52	**0.65**	0.10	**0.69**	**0.55**	**0.74**	**0.68**	0.45	**0.59**	0.04	**0.68**	0.50
Random classifier	-	-	-	-	-	-	-	0.15	0.06	0.03	0.03	0.02	0.07	0.06
CCDb-HG → KTH-Idiap														
Head-Pose	0.44	0.36	0.35	0.72	0.15	0.50	0.40	0.47	0.32	0.29	0.64	0.15	0.49	0.37
Landmarks	0.64	0.41	0.31	0.74	0.06	0.62	0.43	0.65	0.40	0.26	0.65	0.05	0.58	0.40
Head-Pose + Landmarks	0.64	0.47	**0.38**	0.75	0.06	0.63	0.46	**0.66**	0.43	**0.34**	0.67	0.06	0.60	0.43
Head-Pose + **Gaze**	0.53	0.22	0.29	0.76	**0.28**	0.52	0.41	0.52	0.23	0.22	0.66	**0.19**	0.48	0.36
Landmarks + **Gaze**	**0.67**	0.51	**0.38**	**0.80**	0.11	**0.67**	**0.49**	0.64	0.43	0.28	**0.70**	0.10	**0.61**	0.43
Landmarks + Head-Pose + **Gaze**	0.63	**0.54**	**0.38**	0.78	0.06	0.65	0.48	0.64	**0.47**	0.32	0.69	0.05	**0.61**	**0.44**
Random classifier	-	-	-	-	-	-	-	0.12	0.03	0.02	0.10	0.02	0.07	0.06

TABLE III: Impact of the input modalities when training on the CCDb-HG dataset. Invariance is employed as input representation. Results when training on KTH-Idiap dataset are in appendix.

	I^m	Event-based F1		Frame-based F1	
		KTH-Idiap	CCDb-HG	KTH-Idiap	CCDb-HG
Trained on CCDb-HG					
Baseline	Abs	0.59	0.68	0.56	**0.68**
Baseline + Geom-DA	Abs	0.63	**0.69**	0.60	**0.68**
Baseline	Rel	0.57	**0.69**	0.51	0.67
Baseline + Geom-DA	Rel	**0.65**	0.67	0.60	**0.68**
Baseline + Invariance	Inv	**0.65**	**0.69**	**0.61**	**0.68**
Trained on KTH-Idiap					
Baseline	Abs	0.62	0.48	0.60	0.48
Baseline + Geom-DA	Abs	0.61	0.56	0.61	0.55
Baseline	Rel	0.67	0.56	**0.65**	0.56
Baseline + Geom-DA	Rel	0.67	**0.60**	0.64	**0.61**
Baseline + Invariance	Inv	**0.68**	**0.60**	**0.65**	0.59

TABLE IV: Viewpoint invariance (Geom-DA, Invariance).

	Event-based F1		Frame-based F1	
	KTH-Idiap	CCDb-HG	KTH-Idiap	CCDb-HG
Trained on CCDb-HG				
Invariance	**0.65**	0.69	**0.61**	**0.68**
Invariance + TS-DA	0.63	**0.70**	0.59	**0.68**
Invariance + Mixup-DA	0.64	0.69	0.60	0.67
Invariance + TS-DA + Mixup-DA	0.64	0.68	**0.61**	0.66
Trained on KTH-Idiap				
Invariance	**0.68**	0.60	0.65	0.59
Invariance + TS-DA	0.67	**0.62**	**0.66**	**0.61**
Invariance + Mixup-DA	0.67	0.55	0.64	0.51
Invariance + TS-DA + Mixup-DA	**0.68**	0.57	0.65	0.54

TABLE V: Results of temporal data augmentation.

C. Results

Effectiveness of different input cues and gaze as an auxiliary cue. Results based on various input modalities are given in Tab. III. They show that landmarks outperform head-pose, exhibiting a significant 24% relative increase in micro-events F1 and a 10% in macro-events F1, observed in both within- and cross-dataset evaluations. Similar findings are obtained for models trained on KTH-Idiap (see Tab. C in sup. materials). This superiority is attributed to the accuracy of landmarks in providing implicit orientation as well as fine grained dynamic head motion signals.

Regarding gaze, we can notice that using all modalities improves generalization performance from CCDb-HG to KTH-Idiap, with a 4% relative gain in micro- and macro-event F1 compared to using only landmarks and head-pose. More generally, we can see that the addition of gaze to any other modality combination consistently enhances macro-event metrics, with a relative increase of up to 4% within and up to 13% in cross-dataset experiments. Per-class analysis reveals that the benefits of gaze lie mainly in improving Turn and Shake gesture recognitionfor models trained on CCDb-HG, and Turn and Up/Down for models trained on KTH-Idiap (see Table C in sup. materials), whereas the performance on nods and tilt tend to remain the same. which can be due to the distinctive role that gaze plays in these gestures. For instance, we can note from the event confusion matrix (Fig. B in sup. materials) on CCDb-HG that the gaze cue allows to disambiguate between Turn and Shake gestures.

Abs and Rel input representation. The baseline in Tab. IV sheds light on the need for relative (Rel) input representation vs simpler absolute measures (Abs). In low-data regimes like KTH-Idiap, Rel features are more effective, confirming results obtained in previous studies favoring Rel variants [20]–[22], [24], [25], [27]. However, when trained on CCDb-HG, the performance are similar when tested on CCDb-HG, and surprisingly even slightly better on KTH-Idiap despite the larger variability in head orientation compared to the CCDb-HG training data. Additionally, combined with Geom-DA technique, the Abs representation yiels comparable performance to Rel + Geom-DA, suggesting that using absolute representation measures could potentially be sufficient when having enough training data, removing the need for designing specialized features like dynamics.

Viewpoint invariance. Tab. IV further provides a comparative analysis of two methods for achieving viewpoint invariance, either through data augmentation (Geom-DA), or through invariant feature computation (Invariance, *Inv*). First, the Table highlights the sensitivity of simple Abs and Rel features to viewpoint changes. Secondly, it shows that both Geom-DA and Invariance methods significantly enhance generalization performance in cross-dataset settings, with a relative increase of up to 14% and 20% in event-based and frame-based scenarios, respectively. Furthermore, it shows in both cases that models trained on KTH-Idiap and tested on CCDb-HG exhibit a significant performance drop compared to models trained and tested on CCDb-HG (0.60 vs round 0.69 event-F1), whereas models trained on CCDb-HG and tested on KTH-Idiap achieve an F1 score of 0.65, closely matching the performance of models trained and tested on KTH-Idiap (0.68). This highlights the stronger generalization capacity of our larger scale CCDb-HG annotation.

Trained on CCDb-HG	Micro Event F1		Micro Frame F1		Macro Event F1		Macro Frame F1	
	KTH-Idiap	CCDb-HG	KTH-Idiap	CCDb-HG	KTH-Idiap	CCDb-HG	KTH-Idiap	CCDb-HG
Paggio et al. [25]†	0.32 (0.011)	0.50 (0.008)	0.35 (0.011)	0.53 (0.006)	0.15 (0.004)	0.22 (0.004)	0.16 (0.002)	0.21 (0.003)
Otsuka et al. [24]†	0.47 (0.006)	0.53 (0.007)	0.47 (0.007)	0.52 (0.002)	0.35 (0.009)	0.45 (0.013)	0.31 (0.005)	0.41 (0.006)
GRU + Gaze + Invariance	0.58 (0.002)	0.69 (0.005)	0.55 (0.005)	0.68 (0.006)	0.42 (0.015)	0.55 (0.007)	0.38 (0.010)	0.50 (0.004)
LSTM + Gaze + Invariance	0.57 (0.009)	0.69 (0.009)	0.55 (0.008)	0.69 (0.007)	0.42 (0.017)	0.57 (0.021)	0.39 (0.013)	0.52 (0.014)
TCN + Gaze + Invariance	0.61 (0.007)	**0.71 (0.005)**	0.58 (0.005)	**0.71 (0.003)**	0.47 (0.011)	**0.61 (0.005)**	0.42 (0.007)	**0.56 (0.003)**
CNN + Gaze + Invariance	**0.65 (0.018)**	0.69 (0.011)	**0.60 (0.010)**	0.68 (0.008)	**0.48 (0.020)**	0.56 (0.010)	**0.43 (0.009)**	0.51 (0.009)

TABLE VI: Comparison with state-of-the-art methods and investigation of various deep learning models trained on CCDb-HG, including both within-dataset and cross-dataset evaluations. † Our re-implementation. Results are the average over 5 runs and standard deviation are given in parenthesis.

Trained on CCDb-HG	Gaze	Invariance	Micro Event F1		Micro Frame F1		Macro Event F1		Macro Frame F1	
			KTH-Idiap	CCDb-HG	KTH-Idiap	CCDb-HG	KTH-Idiap	CCDb-HG	KTH-Idiap	CCDb-HG
CNN	✗	✗	0.53 (0.024)	0.67 (0.005)	0.50 (0.013)	0.67 (0.003)	0.38 (0.014)	0.52 (0.006)	0.34 (0.010)	0.48 (0.006)
CNN	✓	✗	0.59 (0.017)	**0.69 (0.007)**	0.54 (0.009)	0.67 (0.007)	0.42 (0.023)	0.55 (0.015)	0.37 (0.013)	0.50 (0.010)
CNN	✗	✓	0.63 (0.006)	0.68 (0.002)	0.59 (0.005)	0.67 (0.003)	0.46 (0.010)	0.53 (0.010)	0.42 (0.004)	0.49 (0.008)
CNN	✓	✓	**0.65** (0.018)	**0.69 (0.011)**	**0.60** (0.010)	**0.68 (0.008)**	**0.48** (0.020)	**0.56 (0.010)**	**0.43** (0.009)	**0.51 (0.009)**

TABLE VII: Ablation study evaluating the impact of using *Gaze* as an auxiliary modality and using viewpoint invariance representation. The baseline is *Head-Pose* + *Landmarks*, and in the absence of Invariance, the representation defaults to Rel. Results are the average over 5 runs and standard deviation are given in parenthesis.

Temporal Data augmentation. Tab. V shows empirical findings on using TS-DA and Mixup-DA temporal augmentation techniques for robustness enhancement. Unfortunately, we see that these techniques do not increase recognition, potentially indicating that on our application and datasets, adding temporal noise and fluctuations does not help, either because such noise might already be present in the data, or because it forces the model to detect unwanted signals, leading to the detection of false positives in the real samples.

State of the art (SoA) and recognition models. Table VI, presents a comparison with SoA methods and investigates the effectiveness of various models, including RNN-based architectures (GRU, and LSTM), as well as CNN-based architectures (CNN, and TCN). We can note that our models consistently beats the SoA performance. This outstanding performance primarily originates from the use of additional modalities beyond *Head-Pose*, such as *Landmarks* and *Gaze*, improved input representation (e.g., Invariance), and higher capacity models.

Regarding models, CNN-based architectures outperform RNN-based ones. More specifically, CNN and TCN have superior cross-dataset generalization performance on KTH-Idiap, e.g. with Micro-F1 scores of 0.65 and 0.61 respectively, compared to 0.58 and 0.57 for GRU and LSTM. In within CCDb-HG dataset evaluation, RNN architectures are closer in performance, as these model may benefit more from the larger amount of training data. Nevertheless TCN exhibits superior performance over other models, achieving event-based F1 scores of 0.71 (Micro) and 0.61 (Macro), potentially due to its better ability at modeling longer-range temporal dependencies.

Ablation study. The ablation study in Table VII, evaluates the key contributions of our work: the incorporation of gaze as an auxiliary modality and the use of viewpoint invariance representation. When evaluated on CCDb-HG, introducing gaze results in a statistically significant relative increase of $4\%^{**}$ in macro F1 and $3\%^{*}$ in micro event F1. Cross-evaluation on KTH yields similar results (relative increase ranging from $8\%^{**}$ to $13\%^{**}$). While Invariance does not show improvement within the dataset, as expected, it contributes to enhanced generalization in cross-dataset conditions, exhibiting a significant relative increase from $18\%^{**}$ to $23\%^{**}$. Simultaneously considering both components results in a slight boost in overall performance. This underscores the advantages of gaze and Invariance in enhancing robust HGR across diverse recording and evaluation scenarios.

Limitations. Our proposed method fails to recognize very subtle gestures and out-of-distribution dynamics like fast nodding. Furthermore, the continuous nature of the head pose creates borderline cases such as tilt/nod happening simultaneously, this challenge arises both in annotation and modeling. Body movements, facial expressions, talking, and laughing are also sources of false positives emphasizing the need for further exploration. See supplementary material for a comprehensive discussion and visual examples.

VI. Conclusion

In this work, we introduced CCDb-HG, a novel annotation extension for the CCDb conversational dataset, enriching it with diverse head gesture classes and showing that due to its scale, trained model on CCDb-HG offers superior generalization performance compared to models trained with smaller datasets. We also conducted investigations into the most effective input cues, uncovering the positive impact of gaze as an auxiliary cue, particularly in disambiguating specific gestures and improving representation robustness. We proposed two distinct approaches for achieving viewpoint invariance, showcasing Invariance as an effective approach. Furthermore, while previous research mainly favored relative input representations, we provided evidence that absolute measures, paired with proper data augmentation techniques like Geom-DA and leveraging large-scale datasets like CCDb-HG offer comparable results, reducing the need for hand-crafted features. Finally, we explored various recognition models to establish baseline performance on CCDb-HG, contributing valuable insights regarding head gesture recognition. By releasing our annotations, evaluation code,

and models, we aim to foster reproducibility and stimulate further advancements in non-verbal behavior analysis, for instance to explore other cue fusion strategies or by investigating end-to-end models relying on streams of face images.

REFERENCES

[1] A. J. Aubrey, D. Marshall, P. L. Rosin, J. Vendeventer, D. W. Cunningham, and C. Wallraven. Cardiff conversation database (ccdb): A database of natural dyadic conversations. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 277–282, 2013. 1, 2, 3

[2] Y. Chen, Y. Yu, and J.-M. Odobez. Head nod detection from a full 3d model. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pages 136–144, 2015. 1, 2, 3, 4, 5, 6, 12

[3] J. Chung, C. Gulcehre, K. Cho, and Y. Bengio. Empirical evaluation of gated recurrent neural networks on sequence modeling. *arXiv preprint arXiv:1412.3555*, 2014. 6

[4] Y. A. Farha and J. Gall. Ms-tcn: Multi-stage temporal convolutional network for action segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3575–3584, 2019. 6, 13

[5] D. Gatica-Perez. Automatic nonverbal analysis of social interaction in small groups: A review. *Image and vision computing*, 27(12):1775–1787, 2009. 1

[6] S. Hochreiter and J. Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997. 6

[7] Y. N. D. D. L.-P. Hongyi Zhang, Moustapha Cisse. mixup: Beyond empirical risk minimization. *International Conference on Learning Representations*, 2018. 2, 3, 6

[8] H. Ismail Fawaz, G. Forestier, J. Weber, L. Idoumghar, and P.-A. Muller. Deep learning for time series classification: a review. *Data mining and knowledge discovery*, 33(4):917–963, 2019. 6

[9] D. Jayagopi, D. Sanchez-Cortes, K. Otsuka, J. Yamato, and D. Gatica-Perez. Linking speaking and looking behavior patterns with group composition, perception, and performance. In *Proceedings of the 14th ACM international conference on Multimodal interaction*, pages 433–440, 2012. 1

[10] M. Jensen. Personality traits and nonverbal communication patterns. *Int'l J. Soc. Sci. Stud.*, 4:57, 2016. 1

[11] A. Kapoor and R. W. Picard. A real-time head nod and shake detector. In *Proceedings of the 2001 workshop on Perceptive user interfaces*, pages 1–5, 2001. 2, 3

[12] S. Kawato and J. Ohya. Real-time detection of nodding and head-shaking by directly detecting and tracking the" between-eyes". In *Proceedings Fourth IEEE International Conference on Automatic Face and Gesture Recognition (Cat. No. PR00580)*, pages 40–45. IEEE, 2000. 2, 3

[13] K. Khan, R. U. Khan, R. Leonardi, P. Migliorati, and S. Benini. Head pose estimation: A survey of the last ten years. *Signal Processing: Image Communication*, 99:116479, 2021. 2

[14] S. Kousidis, Z. Malisz, P. Wagner, and D. Schlangen. Exploring annotation of head gesture forms in spontaneous human interaction. In *Proceedings of the Tilburg Gesture Meeting (TiGeR 2013)*, 2013. 2, 3

[15] C. Lea, M. D. Flynn, R. Vidal, A. Reiter, and G. D. Hager. Temporal convolutional networks for action segmentation and detection. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1003–1012, 2017. 2

[16] C. Lea, M. D. Flynn, R. Vidal, A. Reiter, and G. D. Hager. Temporal convolutional networks for action segmentation and detection. In *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 156–165, 2017. 6

[17] C. Lea, R. Vidal, A. Reiter, and G. D. Hager. Temporal convolutional networks: A unified approach to action segmentation. In *Computer Vision–ECCV 2016 Workshops: Amsterdam, The Netherlands, October 8-10 and 15-16, 2016, Proceedings, Part III 14*, pages 47–54. Springer, 2016. 6

[18] P. Lu, M. Zhang, X. Zhu, and Y. Wang. Head nod and shake recognition based on multi-view model and hidden markov model. In *International Conference on Computer Graphics, Imaging and Visualization (CGIV'05)*, pages 61–64. IEEE, 2005. 2, 3

[19] C. Lugaresi, J. Tang, H. Nash, C. McClanahan, E. Uboweja, M. Hays, F. Zhang, C.-L. Chang, M. Yong, J. Lee, W.-T. Chang, W. Hua, M. Georg, and M. Grundmann. Mediapipe: A framework for perceiving and processing reality. In *Third Workshop on Computer Vision for AR/VR at IEEE Computer Vision and Pattern Recognition (CVPR)*, 2019. 4

[20] L.-P. Morency, A. Quattoni, and T. Darrell. Latent-dynamic discriminative models for continuous gesture recognition. In *2007 IEEE conference on computer vision and pattern recognition*, pages 1–8. IEEE, 2007. 2, 7

[21] L.-P. Morency, C. Sidner, C. Lee, and T. Darrell. Contextual recognition of head gestures. In *Proceedings of the 7th international conference on Multimodal interfaces*, pages 18–24, 2005. 2, 7

[22] L. Nguyen, J.-M. Odobez, and D. Gatica-Perez. Using self-context for multimodal detection of head nods in face-to-face interactions. In *Proceedings of the 14th ACM international conference on Multimodal interaction*, pages 289–292, 2012. 2, 3, 7

[23] C. Oertel, K. A. Funes Mora, S. Sheikhi, J.-M. Odobez, and J. Gustafson. Who will get the grant? a multimodal corpus for the analysis of conversational behaviours in group interviews. In *Proceedings of the 2014 Workshop on Understanding and Modeling Multiparty, Multimodal Interactions*, pages 27–32, 2014. 1, 3, 6

[24] K. Otsuka and M. Tsumori. Analyzing multifunctionality of head movements in face-to-face conversations using deep convolutional neural networks. *IEEE Access*, 8:217169–217195, 2020. 1, 2, 3, 7, 8, 11

[25] P. Paggio, M. Agirrezabal, B. Jongejan, and C. Navarretta. Automatic detection and classification of head movements in face-to-face conversations. In *Proceedings of LREC2020 Workshop" People in language, vision and the mind"(ONION2020)*, pages 15–21, 2020. 2, 3, 7, 8, 11

[26] P. Paggio and C. Navarretta. The danish nomco corpus: multimodal interaction in first acquaintance conversations. *Language Resources and Evaluation*, 51:463–494, 2017. 1, 2, 3

[27] M. Sharma, D. Ahmetovic, L. A. Jeni, and K. M. Kitani. Recognizing visual signatures of spontaneous head gestures. In *IEEE Winter Conf. on Applications of Computer Vision (WACV)*, pages 400–408, 2018. 2, 3, 4, 6, 7

[28] D. Smirnov and E. M. Nguifo. Time series classification with recurrent neural networks. *Advanced analytics and learning on temporal data*, 8, 2018. 6

[29] W. Tan and G. Rong. A real-time head nod and shake detector using hmms. *Expert Systems with Applications*, 25(3):461–466, 2003. 2, 3

[30] T. T. Um, F. M. J. Pfister, D. Pichler, S. Endo, M. Lang, S. Hirche, U. Fietzek, and D. Kulić. Data augmentation of wearable sensor data for parkinson's disease monitoring using convolutional neural networks. In *Proceedings of the 19th ACM International Conference on Multimodal Interaction*, ICMI 2017, pages 216–220. ACM, 2017. 2, 3, 6

[31] Z. Wang, W. Yan, and T. Oates. Time series classification from scratch with deep neural networks: A strong baseline. In *2017 International joint conference on neural networks (IJCNN)*, pages 1578–1585. IEEE, 2017. 6, 13

[32] M. Włodarczak, H. Buschmeier, Z. Malisz, S. Kopp, and P. Wagner. Listener head gestures and verbal feedback expressions in a distraction task. In *Feedback Behaviors in Dialog*, 2012. 2

[33] Y. Wu, K. Akiyama, K. Kitani, L. Jeni, and Y. Mukaigawa. Head gesture recognition in spontaneous human conversations: A benchmark. In *Workshop on Egocentric (First-Person) Vision (CVPR)*, volume 2016, page 4, 2016. 1, 2, 3

[34] X. Zhang, S. Park, T. Beeler, D. Bradley, S. Tang, and O. Hilliges. Eth-xgaze: A large scale dataset for gaze estimation under extreme head pose and gaze variation. In *European Conference on Computer Vision (ECCV)*, 2020. 4

[35] X. Zhang, Y. Sugano, and A. Bulling. Revisiting data normalization for appearance-based gaze estimation. In *Proceedings of the 2018 ACM Symposium on Eye Tracking Research amp; Applications*, ETRA '18, 2018. 4

Supplementary Materials

I. Further details of annotations

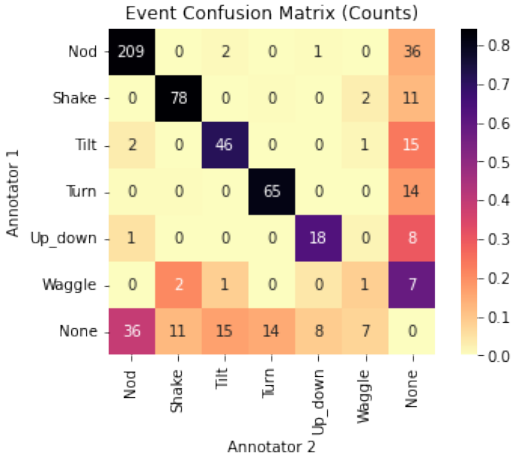
Two annotators independently labeled videos from the CCDb dataset using specified head gesture categories. Each annotator worked on 50 videos and evaluated 15 common videos to gauge inter-annotator agreement. Analysis of the inter-annotator confusion matrix in Fig. 1 indicated disagreements primarily in subtle instances mislead with None, not between different gestures. However, the Waggle category showed notable confusion with Shake, Tilt, and None, at events and frame. Due to low agreement and potential ambiguity, we opted to exclude the Waggle category from subsequent experiments, acknowledging its limited representation in the dataset. Otherwise, we can see that the agreement is relatively high among the other classes, in Fig. 1b, we can see that Up/Down might be misleading with nod, which is logical since the only difference is the gaze behavior.

II. Additional Experiments and Results

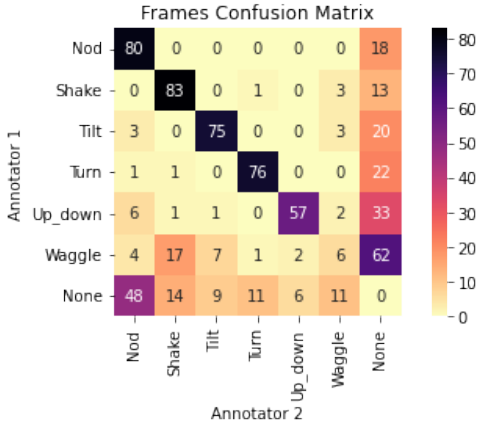
Ablation study. Tab. I presents the ablation study conducted on KTH-Idiap to further assess the primary contributions of this work, which include the incorporation of gaze as an auxiliary modality and the utilization of Invariance as the preferred approach for achieving viewpoint invariance in HGR. When considered independently, they both improve performance, but the results clearly indicate that the most optimal performances are achieved when both gaze and Invariance are employed simultaneously. Specifically, in cross-dataset settings, we observe notable relative increases of up to 7% at the Micro-level and 15% at the Macro-level when trained on KTH-Idiap and evaluated on CCDb-HG.

Comparison to state of the art on KTH-Idiap. Tab. II provides an extension of the evaluation for various models trained on KTH-Idiap, yielding similar observations as in Tab.VI of the main submission. Notably, our results consistently surpass state-of-the-art models, showcasing the positive impact of incorporating gaze and leveraging Invariance. While LSTM exhibits comparatively lower performance, the results for TCN and CNN are closely aligned. In our findings, it becomes challenging to distinctly favor one over the other; when one model outperforms in a specific metric, the other excels elsewhere. When trained on KTH-Idiap, both TCN and CNN demonstrate comparable and potentially suitable performances for HGR.

Confusion matrices: gaze vs. w/o gaze. To further quantify the impact of gaze, we compute the difference between the confusion matrices of the model trained with all cues (*Head-Pose + Landmarks + Gaze*) and the same model trained with *Head-Pose + Landmarks* cues, as presented in Fig. 2 and Fig. 3. In this analysis, positive values along the diagonal and negative off-diagonal values indicate gaze's positive contributions. From examining the within CCDb-HG



(a) Event confusion matrix corresponding to the average between two annotators. Values correspond to the count and the color to the percentage normalized by row.



(b) Frame confusion matrix corresponding to the average between two annotators. Values and colors correspond to the percentage normalized by row. Note that we remove the number of None-None before normalization for visualization purposes.

Fig. 1: Event and frame confusion matrix between two annotators over 15 randomly selected videos.

differences [1] (Fig. 2), it is evident that gaze significantly aids in disambiguating Turn and Shake gestures. However, there is a smaller degradation in confusion related to Nod with Up/Down and None, as well as Tilt. For cross-dataset evaluation from CCDb-HG to KTH-Idiap (Fig. 3), the presence of positive values on the diagonal as well as negative values in the "None" column suggests that incorporating gaze results in improving accuracy and fewer false positives.

Effectiveness of different input cues and gaze as an auxiliary cue, on KTH-Idiap dataset. Tab. III presents the investigation results of various input cues and gaze as an auxiliary cue on KTH-Idiap. Consistent with main submission findings, landmarks significantly outperform head-pose, with up to 9% and 23% relative increases in within-dataset and

[1] Note that the raw differences must be compared to the amount of observed events, which have different orders of magnitude (Nod: 535, Shake: 159, Tilt: 123, Turn: 126, Up/Down: 62 in the CCDb-HG test set, and Nod: 194, Shake: 42, Turn: 216, Tilt: 29, Up/Down: 32 in KTH-Idiap).

Trained on KTH-Idiap	Gaze	Invariance	Micro Event F1		Micro Frame F1		Macro Event F1		Macro Frame F1	
			KTH-Idiap	CCDb-HG	KTH-Idiap	CCDb-HG	KTH-Idiap	CCDb-HG	KTH-Idiap	CCDb-HG
CNN	✗	✗	0.66	0.56	0.63	0.57	0.47	0.40	0.42	0.35
CNN	✓	✗	0.67	0.56	**0.65**	0.56	0.51	0.39	**0.45**	0.33
CNN	✗	✓	0.65	0.59	0.63	**0.59**	0.48	0.43	0.41	0.38
CNN	✓	✓	**0.68**	**0.60**	**0.65**	**0.59**	**0.51**	**0.46**	**0.45**	**0.41**

TABLE I: Ablation study showcasing the impact of incorporating *Gaze* as an auxiliary modality and using Invariance for viewpoint invariance in KTH-Idiap. The baseline modality (without gaze) is *Head-Pose* + *Landmarks*, and in the absence of Invariance, the representation defaults to Rel.

Trained on KTH-Idiap	Micro Event F1		Micro Frame F1		Macro Event F1		Macro Frame F1	
	KTH-Idiap	CCDb-HG	KTH-Idiap	CCDb-HG	KTH-Idiap	CCDb-HG	KTH-Idiap	CCDb-HG
Paggio et al. [25]†	0.42	0.36	0.32	0.33	0.19	0.15	0.15	0.13
Otsuka et al. [24]†	0.59	0.48	0.56	0.44	0.42	0.40	0.38	0.33
LSTM + Gaze + Invariance	0.60	0.56	0.60	0.54	0.44	0.41	0.38	0.36
TCN + Gaze + Invariance	0.67	**0.61**	0.64	**0.61**	**0.52**	**0.48**	**0.46**	**0.42**
CNN + Gaze + Invariance	**0.68**	0.60	**0.65**	0.59	0.51	0.46	0.45	0.41

TABLE II: Comparison with state-of-the-art methods and investigation of various deep learning models for HGR on KTH-Idiap datasets, including both within-dataset and cross-dataset evaluations. † Our re-implementation.

Input Representation	Event-based F1							Frame-based F1						
	Nod	Shake	Tilt	Turn	Up-Down	All-Micro	All-Macro	Nod	Shake	Tilt	Turn	Up-Down	All-Micro	All-Macro
KTH-Idiap														
Head-Pose	0.58	0.34	**0.36**	0.76	0.19	0.59	0.44	0.59	0.28	0.27	0.67	0.13	0.57	0.39
Landmarks	**0.72**	0.36	0.29	0.74	0.11	0.64	0.45	**0.73**	0.29	0.23	0.65	0.09	0.62	0.40
Head-Pose + Landmarks	0.71	0.40	**0.36**	0.76	0.18	0.65	0.48	**0.73**	0.30	0.23	0.66	0.13	0.63	0.41
Head-Pose + **Gaze**	0.54	0.34	0.30	0.78	0.17	0.58	0.42	0.55	0.27	0.23	**0.69**	0.14	0.55	0.38
Landmarks + **Gaze**	0.71	0.37	0.34	0.79	0.31	0.67	0.50	**0.73**	0.29	**0.28**	**0.69**	0.24	**0.65**	**0.45**
Landmarks + Head-Pose + **Gaze**	0.71	**0.41**	0.30	**0.80**	**0.32**	**0.68**	**0.51**	**0.73**	**0.31**	0.22	**0.69**	**0.29**	**0.65**	**0.45**
KTH-Idiap → CCDb-HG														
Head-Pose	0.55	0.58	0.47	0.44	0.15	0.51	0.44	0.51	0.54	0.41	0.41	0.10	0.48	0.39
Landmarks	0.72	0.52	0.27	0.48	0.15	0.59	0.43	0.69	0.44	0.24	0.39	0.09	**0.59**	0.37
Head-Pose + Landmarks	**0.73**	**0.57**	0.28	0.44	0.14	0.59	0.43	**0.70**	**0.49**	0.25	0.38	0.09	**0.59**	0.38
Head-Pose + **Gaze**	0.56	0.43	**0.35**	**0.57**	0.27	0.51	0.43	0.48	0.34	0.28	**0.49**	0.20	0.44	0.36
Landmarks + **Gaze**	0.72	0.50	0.29	0.53	**0.30**	**0.60**	**0.47**	0.67	0.37	0.25	0.47	**0.21**	0.57	0.39
Landmarks + Head-Pose + **Gaze**	0.72	0.53	0.31	0.50	0.24	**0.60**	0.46	0.69	0.44	**0.29**	0.46	0.18	**0.59**	**0.41**

TABLE III: Exploration of diverse combinations of input modalities, incorporating gaze as an auxiliary modality, on the KTH-Idiap dataset. The Invariance variant is employed for input representation.

cross-dataset evaluations, respectively. Using all modalities enhances generalization performance, showing an 8% relative gain from KTH-Idiap to CCDb-HG compared to using only landmarks and head-pose, leading to a robust representation of head dynamics. Adding gaze to Landmarks and Head-Pose + Landmarks also consistently improves macro-event metrics, with up to 13% and 9% relative increase in within and cross-dataset evaluations, respectively. Gaze notably aids in recognizing Turn and Up/Down categories within the KTH-Idiap dataset.

Qualitative results of gaze inclusion. The efficacy of gaze in enhancing recognition was shown through its ability to disambiguate between turn and shake gestures. This point is illustrated in Fig. 6. In the first example (first two rows), it is evident that gaze played a crucial role in distinguishing between a turn and a shake. The gaze remained fixed during side-to-side head movement, characteristic of a shake, as opposed to a turn. Conversely, in the second example (last two rows), a side head movement accompanied by a gaze shift indicated a turn amidst two shake gestures. Once again, the incorporation of gaze proved instrumental in disambiguating these distinct gestures.

Limitations. Visual inspection of model prediction on the test set exposes certain limitations in the model's performance. Notably, we observe that subtle gestures are hard to recognize by the model which translates to false positives. While increasing the decision threshold reduces the number of false negatives, it also increases the number of false positives impacting the prediction quality. Additionally, the recognition accuracy diminishes for out-of-distribution head gesture dynamics, particularly fast nodding. We notice that the continuous nature of the head pose presents a challenge to discretizing head gestures into distinct classes. Qualitative examples illustrating these shortcomings are presented in Fig. 7. For instance, the first three rows illustrate confusion with a tilt gesture. In each case, there is a sideways rotation along the roll axis indicative of a tilt; however, the model predicts a nod in the first row, a turn in the second row, and a shake in the third row. It is important to note that neither the annotation nor prediction is inherently incorrect, and these instances could be characterized as borderline cases or limitations of the class definition as seen in the inter-annotator agreement in Fig. 1. In the last row of the same figure, four examples of correct predictions are presented where the model's confidence was not sufficient, resulting in shorter predicted durations compared to the actual ground

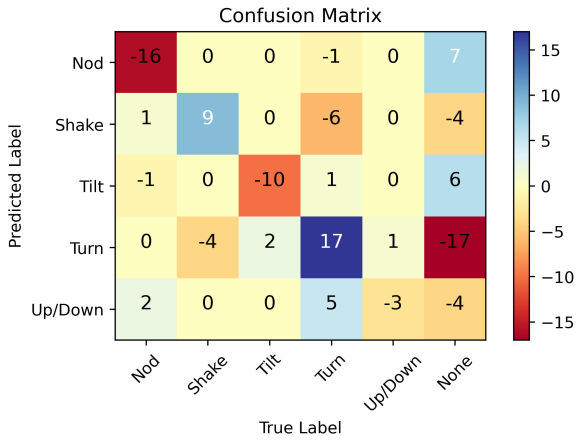


Fig. 2: Difference between the confusion matrix of the model trained with all cues *(Head-Pose+Landmarks+**Gaze**)* and the model trained with *(Head-Pose+Landmarks)* on the CCDB-HG dataset. Positive values on the diagonal, as well as negative values on the off-diagonal, indicate the positive impact of the gaze cue. Similarly, negative values in the "None" column indicate a positive effect due to a reduction of false alarms for the corresponding row.

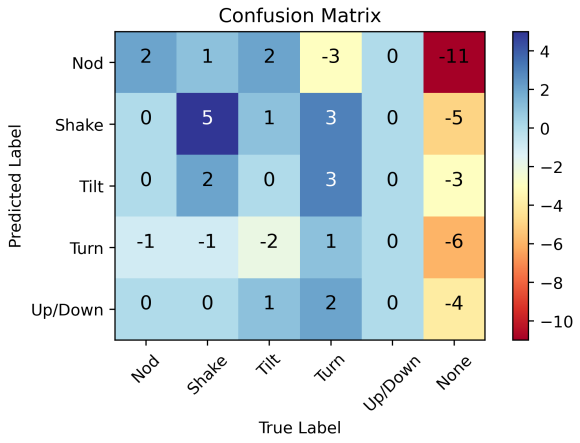


Fig. 3: Difference between the confusion matrix of the model trained with all cues *(Head-Pose+Landmarks+**Gaze**)* and the model trained with *(Head-Pose+Landmarks)* for CCDb-HG $\rightarrow$ KTH-Idiap. Positive values on the diagonal, as well as negative values on the off-diagonal, indicate the positive impact of the gaze cue. Similarly, negative values in the "None" column indicate a positive effect due to a reduction of false alarms for the corresponding row.

truth. In cases of insufficient overlapping, these events are counted as errors. Additionally, false positives may occur when body movement caused by repositioning or laughing induces head motion interpreted by the model as head gestures. Facial movements, such as expressions or talking, can also influence landmark dynamics, leading to false positive predictions.

Camera-relative gaze vs. head-relative gaze. Gaze direction can be represented in different coordinate systems like the camera and head coordinate systems. Although gaze in the head coordinate system is naturally invariant to the viewpoint, making it a theoretically advantageous representation, its effective use requires accurate head pose estimation to convert gaze from camera coordinates. Our empirical comparison in Tab. IV reveals that gaze in camera coordinates outperforms that in head coordinates. This discrepancy may result from the noisy head pose estimator used, affecting the accurate representation of gaze in the head coordinate system.

III. Details of Evaluation Protocol

Evaluation on CCDb-HG is based on a fixed test set including 4 subjects and 25 videos. It avoids repetitive training, allowing future work to use larger end-to-end models. Regarding the nature of KTH-Idiap data, leave-one-person cross-validation is performed for all the experiments. The model is trained with samples from all the videos except from one person and the model is evaluated on all the videos from the excluded person. For each video in the test set, the model is applied to each frame. Following the evaluation protocol from [2], the performance of the model is measured at two different levels:

Frame-based describes well the sensitivity of the model prediction. It uses the standard precision, recall, and F1 score measures.

Event-based describes well the gesture detection capability of the model. Suppose e_i^{gt} is a ground truth event in the time interval I_i^{gt} and e_j^d is a detected event of the same class in the time interval I_j^d. Then, the event matching precision, recall, and F-score between e_i^{gt} and e_j^d is:

$$P_{i,j} = \frac{|I_i^{gt} \cap I_j^d|}{|I_i^{gt}|}, \quad R_{i,j} = \frac{|I_i^{gt} \cap I_j^d|}{|I_j^d|}, \quad F_{i,j} = \frac{2P_{i,j}R_{i,j}}{P_{i,j} + R_{i,j}} \tag{1}$$

Two events match if the F score is above a threshold. In the case of long head gesture, it is hard for the predicted event to be as long as the ground truth thus the threshold is set as 0.1 in order to handle such situations. Finally, given matched events, we can compute event precision, recall, and F-score as follows:

$$P_{event} = \frac{\#\{e_j^d \mid \exists i,\, F_{i,j} > \text{threshold}\}}{\#e^d}$$

$$R_{event} = \frac{\#\{e_i^{gt} \mid \exists j,\, F_{i,j} > \text{threshold}\}}{\#e^{gt}}$$

$$F_{event} = \frac{2P_{event}R_{event}}{P_{event} + R_{event}}$$

Moreover, before computing the evaluation metrics we perform a label smoothing on each predicted video. It consists of a majority vote on a 15 frames window. Thus, it aggregates two similar events that are less than 7 frames apart and it deletes events that are less than 7 frames. Additionally, for the frame-based measure, four frames at the edge of ground truth events are not taken into account. In fact, boundary annotations lack precision thus we don't want this to be reflected in the performance measures.

	Gaze ref	Micro Event F1		Micro Frame F1		Macro Event F1		Macro Frame F1	
		KTH-Idiap	CCDb-HG	KTH-Idiap	CCDb-HG	KTH-Idiap	CCDb-HG	KTH-Idiap	CCDb-HG
Trained on CCDb-HG									
Head-Pose + Landmarks	-	0.63	0.68	0.60	**0.68**	0.46	0.55	0.43	0.50
Head-Pose + Landmarks + Gaze	head	0.63	0.68	0.58	0.67	0.47	**0.56**	0.41	**0.51**
Head-Pose + Landmarks + Gaze	camera	**0.65**	**0.69**	**0.61**	**0.68**	**0.48**	0.55	**0.44**	0.50
Trained on KTH-Idiap									
Head-Pose + Landmarks	-	0.65	0.59	0.63	**0.59**	0.48	0.43	0.41	0.38
Head-Pose + Landmarks + Gaze	head	0.67	0.59	0.63	0.58	0.49	0.43	0.42	0.38
Head-Pose + Landmarks + Gaze	camera	**0.68**	**0.60**	**0.65**	**0.59**	**0.51**	**0.46**	**0.45**	**0.41**

TABLE IV: Ablation study comparing *Gaze* in head coordinates vs. camera coordinates, with *Invariance* as input representation. Baseline modality is *Head-Pose* + *Landmarks* (w/o gaze).

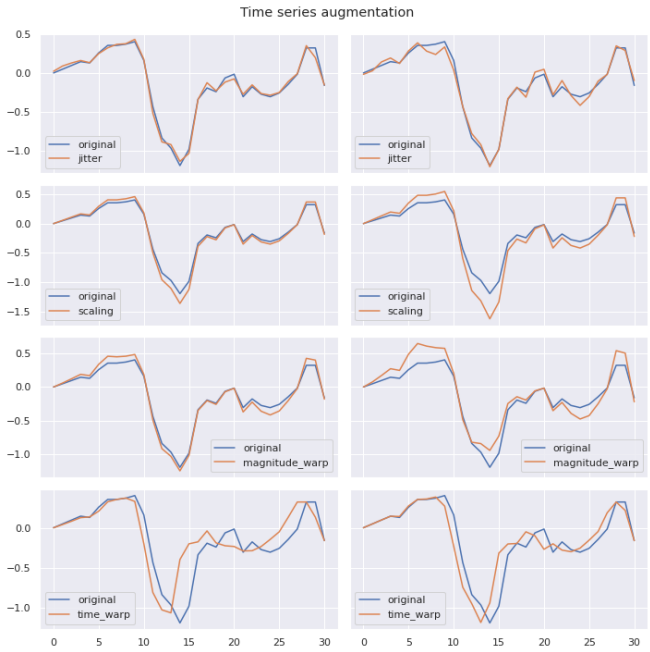


Fig. 4: Examples of the four different time series data augmentation including Jittering, Scaling, Magnitude Warping, and Time Warping applied to a channel of a relative sample.

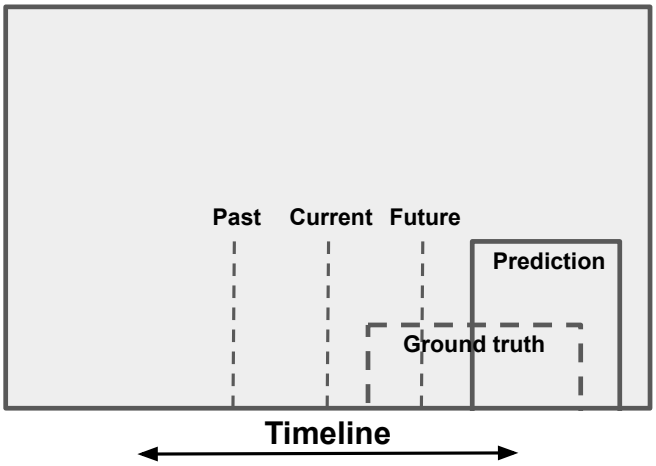


Fig. 5: Diagram presenting the plots in each frame in Fig. 6 and Fig. 7. At the bottom of each frame, color curves represent the event labels over time (120 frames), Nod (Green), Shake (orange), Tilt (light green), Turn (blue), and Up Down (purple). Dash lines are ground truth and solid lines are predictions. The three vertical white dash lines are respectively past, current, and future which represent the input time windows used to predict head gestures.

IV. Further Details of Data Augmentation

In this work, a series of data transformations are applied, encompassing jittering ($p = 0.5$), scaling ($p = 0.66$), magnitude warping ($p = 0.33$), and time warping ($p = 0.3$), where p signifies the probability of applying each transformation. Jittering involves adding random noise sampled from a normal distribution with a mean of zero and a standard deviation (std) of 0.05. Scaling is executed in two mutually exclusive manners: either by multiplying the time-series with a random scalar drawn from a normal distribution with a mean of 0.2 and a std of 1.0, or through magnitude warping utilizing four knots with a std of 0.2. Additionally, time warping is incorporated, involving two knots with a std of 0.1.

V. Details of Models

Each of the implemented models consists of an encoder CNN, LSTM, or TCN. Then, we used an average pool across time followed by a linear head that classifies the head gestures.

CNN. In this work, our configuration closely follows the specifications outlined in [31]. It consists of 3 convolutional layers with kernel sizes ranging from 8,5 and 3 with a stride of 1. We used 128 channels in all three layers. In addition, we used a dropout layer after the two first layers.

RNN. In our LSTM implementation, we adopt a conventional stacked bidirectional LSTM (BiLSTM) architecture, consisting of two layers with a hidden dimension set to 64. At each time step, we extract the hidden representation from the last layer, which is the concatenation of both directions in the LSTM. A similar design for GRU is employed without bidirectional modeling.

TCN. In our implementation of the Temporal Convolutional Network (TCN), we adhere to the Single-Stage TCN model introduced in [4]. Initially, we employ a 1x1 convolutional layer in 1D to project each time step into 128 hidden dimensions. Subsequently, we stack four dilated residual layers, following the structure outlined in [4]. These layers utilize 1D convolutional layers with a kernel size of 3, and the dilation rate increases progressively from 1, 2, 4, to 8.

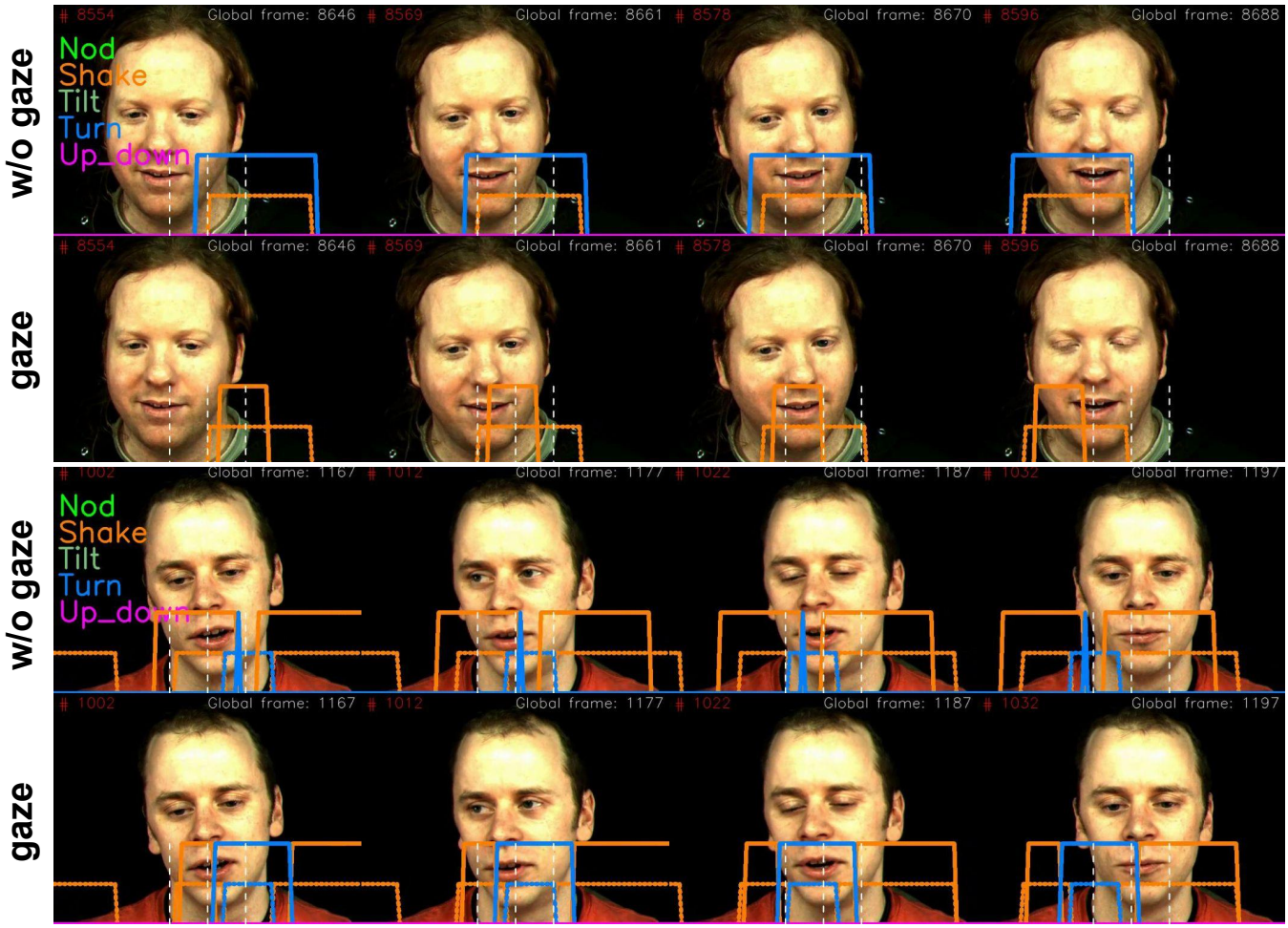


Fig. 6: Comparison of results with and without (w/o) gaze. A diagram in Fig. 5 gives a comprehensive understanding of the plots in each frame. The first two rows show a ground truth shake, where the gaze can disambiguate a turn with a shake. The last two rows focus on a ground truth turn, where including gaze helped to recognize the turn. We can notice that in both cases the gaze is a relevant cue to disambiguate gestures.

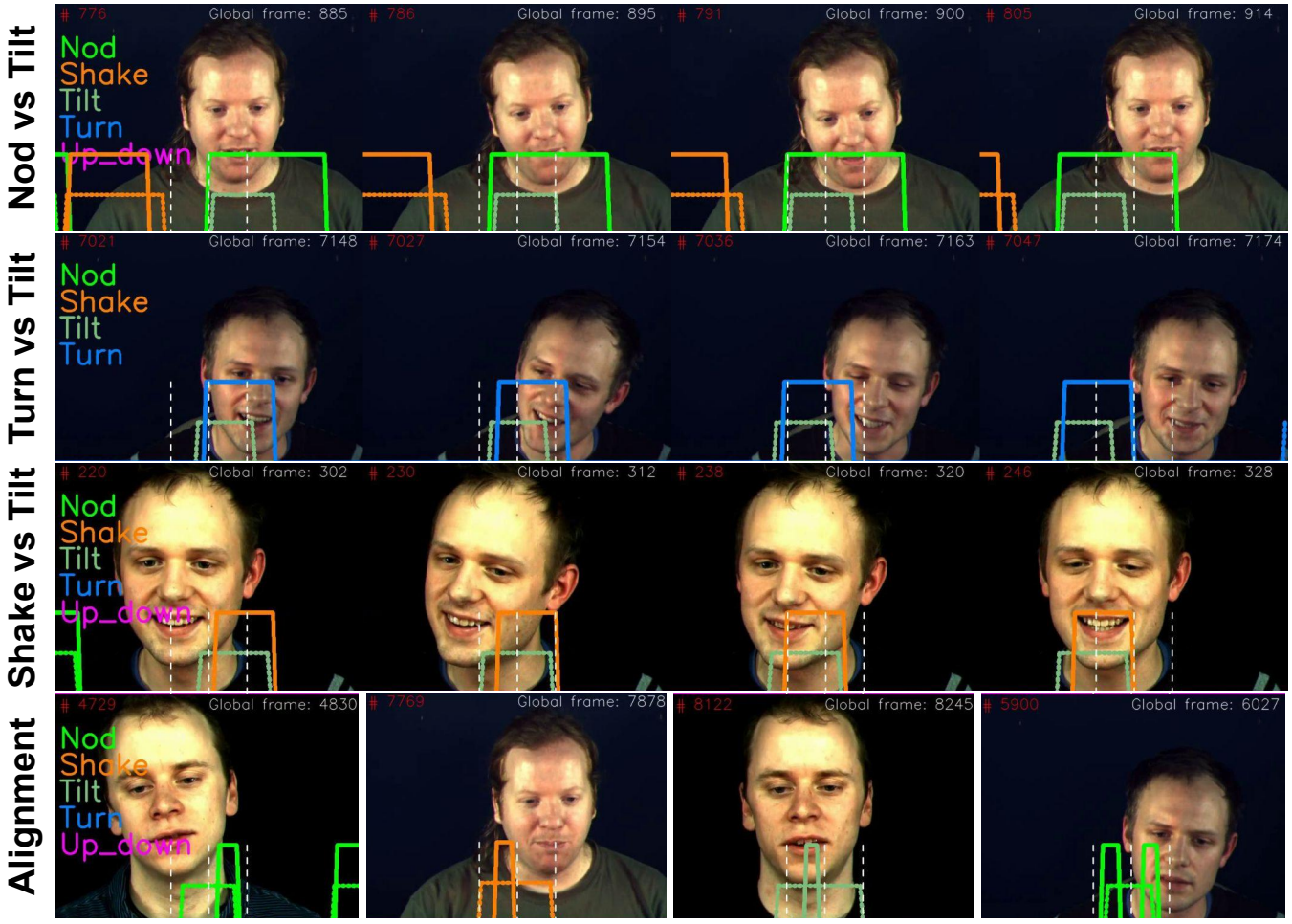


Fig. 7: Error predictions from a CNN model using head pose, landmarks, and gaze. A diagram in Fig. 5 gives a comprehensive understanding of the plots in each frame. The first three rows show predicted events confused with a ground truth tilt. The last row shows four misalignments where prediction events are correct but the overlap with the ground truth is too small.