Head and Body Orientation Estimation with Sparse Weak Labels in Free Standing Conversational Settings

Stephanie Tan S.TAN-1@TUDELFT.NL
David M.J. Tax D.M.J.TAX@TUDELFT.NL
Hayley Hung H.HUNG@TUDELFT.NL
Delft University of Technology

Abstract

We focus on estimating human head and body orientations which are crucial social cues in free-standing conversational settings. Automatic estimations of head and body orientations enable downstream research about conversation involvement, influence, and other social concepts. However, in-the-wild human behavior and long interaction datasets are difficult to collect and expensive to annotate. Our approach mitigates the need for large number of training labels by casting the task into a transductive low-rank matrix-completion problem using sparsely labelled data. We differentiate our learning setting from the typical data-intensive setting required for existing supervised deep learning methods. In situations of low labelled data availability, our method takes advantage of the inherent properties and dynamics of the social scenarios by leveraging different sources of information and physical priors. Our method is (1) data efficient and uses a small number of annotated labels, (2) ensures temporal smoothness in predictions, (3) adheres to human anatomical constraints of head and body orientation differences, and (4) exploits weak labels from multimodal wearable sensors. We benchmark this method on the challenging multimodal SALSA dataset, the only large scale dataset that contains video, proximity sensors and microphone audio data. When only using 5% of all the labels as training samples, we report 65% and 76% averaged classification accuracy for head and body orientations, which is an 8% and 16% respective increase compared to previous state-of-the-art performance under the same transductive setting.

Keywords: Human orientation estimation, matrix completion, weak labels, free-standing conversations

1. Introduction

Studying social scenes that have free-standing conversation groups (FCGs) is of great interest. FCGs are a type of focused encounters that emerge in many social occasions, such as a cocktail party, a coffee break, a networking event, etc (Setti et al., 2015). We find relevance in studying these social entities in order to study human interactions as part of the complex social dynamics. Prominent non-verbal cues that depict the social interplays are participant head and body orientations. With accurate estimations of head and body orientations, high-level social concepts such as conversation group formations and schisms can become more explainable.

Head and body orientations of participants are necessary prerequisites for many downstream tasks such as turn-taking patterns, conversation group memberships, estimation of social attention, etc (Ba and Odobez, 2009). Some tasks may only require either the head

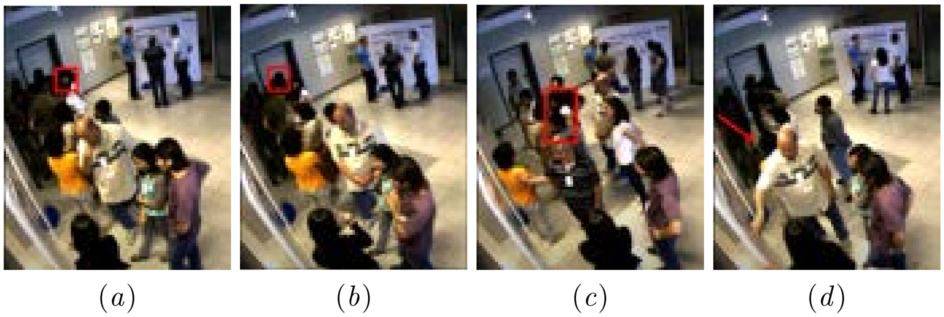


Figure 1: Examples of head and body orientation estimation challenges from the SALSA dataset (Alameda-Pineda et al., 2016) as highlighted in red: (a) low resolution, (b) low visibility, (c) background clutter, and (d) occlusion.

or body orientation. When identifying addresser/addressee or speaker/listener in conversations, head orientations are the primary cues (Huang et al., 2011). When estimating group memberships, body orientations are the primary cues (Kendon, 1990). However, Langton et al. (Langton et al., 2000) have shown that head and body orientations are both important cues for estimating social attention. In social scenes such as Figure 1, eye gaze direction cannot be reliably observed; the attention target positions are not fixed throughout time, and the number of attention targets is not predefined. Under these adverse circumstances, attention direction is difficult to estimate. Hence, the importance of robust and accurate head and body orientations becomes more evident.

While there are many successes in human pose estimation and orientation estimation using deep learning frameworks (Güler et al., 2018; Wei et al., 2016; Toshev and Szegedy, 2014; Tompson et al., 2014), these methods only work well when human faces and body parts are easily discernible. Head and body orientation estimation remain challenging, especially for crowded scenes with relatively static subjects captured by videos from elevated side-views which result in low resolution, low light visibility, background clutter and occlusions (Figure 1 for example) (Hu et al., 2004). In these settings, off-the-shelf deep learning methods are not effective (Carissimi et al., 2018) and retraining/finetuning them requires a considerable number of labelled samples. This motivates our proposed method under the transductive and few-labels setting which simultaneously estimates head and body orientations by leveraging wearable sensor data in addition to videos.

Recent advances have shown the efficacy in using a multi-view camera and multi-sensor scenario (Alameda-Pineda et al., 2016; Tan et al., 2021). The multi-view camera setting offers different viewpoints on people in the scene for better acquisitions of head and body orientations. More interestingly, wearable sensors such as inertial measurement units (IMUs), microphones, infrared or Bluetooth proximity sensors, etc. have demonstrated an ability to recover subject orientations independently of the video modality (Canton-Ferrer et al., 2008; Kok et al., 2017; Kok and Schön, 2019). In scenarios where video and microphone audio data are both recorded, a multimodal approach of head orientation estimation can be more accurate and robust than a unimodal one, as shown by Canton-Ferrer et al. (2008). Microphone data indicate who the speaker is at a given moment, and it is well known that the speaker tends to be the center of visual focus of a conversation group (Massé et al., 2018). Considering these two aspects and given the ground positions of the interactants in free-standing scenarios, head orientations can be more reliably predicted in a complementary manner, especially when video data is partial or missing.

Despite the benefits that multimodal data from wearable sensors may offer, it is challenging to work with them. This is illustrated most evidently by the lack of in-the-wild datasets capturing natural interactions and emphasizing ecological validity in this domain, as it requires monumental effort to collect and annotate. Malfunctions of wearable sensors during data collection are more difficult to notice compared to those of video cameras. The types of different sensor noise are also hard to characterize. The resulting data could be of low quality, partial and/or missing due to periodic dropouts in sensor data streams, etc. (Higger et al., 2013; Newman et al., 2018). However, we argue that these difficulties are not reasons to deter from exploiting multimodal data from wearable sensors because the available data could still be of great value, as shown by literature (Alameda-Pineda et al., 2016; Tan et al., 2021).

In this work, we highlight the possibility and advantage of working with a small number of human annotated orientation labels, along with sparse, noisy but automatically acquired labels from wearable sensors. As mentioned previously, wearable sensors are hard to work with. While it is possible to estimate labels from wearable sensors, the label quality varies depending on raw wearable sensor data quantity and quality. Hence we refer them as weak labels in this paper. Our results show that having information provided by other modalities like wearable sensors can indeed improve the performance of head and body orientation estimations in this free-standing conversation setting.

This study simultaneously addresses the following context where: 1) there is a relatively small number of head and body orientation samples ($\sim 10^2 - 10^3$) for each subject, 2) we jointly predict head and body orientation classification labels for unobserved samples only using a very small number ($\sim 5\%$) of sparsely distributed ground truth labels, 3) we take advantage of the temporal structure within the orientation label data and improve upon a previously suggested model based on Gaussian process regression (GPR) (Tan et al., 2018), and 4) most importantly, we fully exploit the utility of head and body orientation weak labels in addition to very few ground truths to improve performance.

2. Related Work

2.1. Human pose estimation

Recent developments of deep learning methods (Cao et al., 2017; Fang et al., 2017; Kreiss et al., 2019; Bazarevsky et al., 2020) had greatly advanced 2D human pose estimation. Even though results are promising, addressing existing challenges such as low resolution and heavily occluded targets (Carissimi et al., 2018), and cluttered and crowded backgrounds, is an active research topic. Popular off-the-shelf pose estimation methods such as Openpose (Cao et al., 2017) use a bottom-up approach to first detect body joints and later form associations to estimate a full skeleton model for each person in the frame. However, having only body part locations does not provide enough information to directly estimate the orientations of those body parts.

Using 3D pose estimation methods or converting 3D pose datasets (Ionescu et al., 2013) allows for extraction of orientations. Recent methods focusing on 3D pose estimations (full body, hand+body, etc.) Rong et al. (2020); Sárándi et al. (2020); Choutas et al. (2020) could be promising to directly infer orientations from 3D skeletal poses. However, orientation estimation could be decoupled and simplifed from 3D pose estimation problem as there is evidence showing orientation estimations for objects can perform better when using 2D image features than 3D landmarks (Ghodrati et al., 2014). 3D poses may be difficult to infer due to occlusion or low resolution body parts, which are relevant scenarios in crowded social interactions in-the-wild. To address occlusion for 3D poses is an ongoing topic with (Veges and Lőrincz, 2020) showing initial success on the MuPoTS dataset through localization and pose estimation with temporal smoothing.

2.2. Head and body orientation estimation: RGB data

Previous works (e.g., Sigal and Black 2006; Ba and Odobez 2009) in head and body orientation estimation saw successes in using methods based on probabilistic frameworks (e.g. dynamic Bayesian networks, hidden Markov models, etc.). Taking advantage of the physical constraint of relative head and body pose and walking direction, Chen et al. (2011) focus on the joint estimation of head and body orientation to achieve improved results. This body of work targets orientation estimations in a specific context by exploiting facial landmarks or motion priors; while this paper differentiates itself by focusing on the task in the surveillance setting with relatively static subjects. Without large movement towards one direction as a cue, orientation estimation becomes more difficult. Overall, there is more previous work on head orientation estimation compared to that of the body in the surveillance and crowded setting. In this particular context, human heads can be more easily seen and therefore head orientations are easier to predict. Human bodies can be occluded, making body orientations predictions more difficult. Lee et al. (2017) proposed CRPNet that works well with low resolution images. However, their design goal favors speed over accuracy.

We acknowledge that there is a number of deep learning based methods (Beyer et al., 2015; Prokudin et al., 2018) for head and/or body orientation estimation problems. Most available methods are trained on datasets (Tosato et al., 2012) that contain facial views. Applying Beyer et al.'s method (Beyer et al., 2015) to SALSA is not straightforward because of the multi-camera setting and the extent of facial and body part occlusions. A body orientation estimation method proposed by Choi et al. (2016) also faces similar challenges as head orientation estimation methods. Raza et al. (2018) reported a joint head and body orientation estimation model using a hierarchical convolutional neural network. This pre-trained model trained with relatively small datasets (e.g., Human3.6M, Ionescu et al. 2013) would most likely only be suitable for estimating orientations for pedestrians, and not for crowded and static social scenes like SALSA. Overall, the development of generalizable deep learning solutions for head and body orientation estimations are held back because of the lack of large scale datasets and the lack of environment/context variety in the training images. This constraint was only recently pointed out and addressed by the release of the COCO-MEBOW dataset (Wu et al., 2020), which would enable future new data-intensive head and body orientation estimation methods.

Previous works (Beyer et al., 2015; Hasan et al., 2018) showed that regression of head orientations can be achieved. Tasks such as predicting social attention (Massé et al., 2017) and personality traits (Subramanian et al., 2013) may benefit from more fine-grained orientation estimations. While regression is more descriptive, it is also challenging compared to orientation classification. As indicated by Tan et al. (2021), the annotation noise for head orientation labels from video annotations is around 17°, which is more than the bin-width of class in our setting. Further experiments show that regression could be more advantageous, but the increase in performance on average is small compared to the variance. For the scope of this work, we reduce the orientation estimation problem to an 8-class classification problem (i.e., dividing 360° into eight sectors).

2.3. Head and body orientation estimation: depth and wearable sensors

Depth images can be used in estimated orientations. However, many works in this area (e.g., Fanelli et al. 2011; Shinmura et al. 2015; Okuno et al. 2018) rely on the detection of the face and/or localization of facial and body landmarks. Works such as Liu et al. (2013) combines RGB and depth data to estimate human body orientations. It is still challenging for subjects in crowded social scenes because of heavy occlusions with little motion cues.

In the sensor signal processing community, it is common practice to use wearable systems that house IMU sensors. To estimate orientations, IMU data serve as inputs to algorithms such as quaternion-based Extended Kalman Filtering and more recently reinforcement learning based methods (Kok et al., 2017; Kok and Schön, 2019; Laidig et al., 2021; Hu et al., 2021). Ahmed and Tahir (2017) showed that errors in estimating body part orientations while doing multi-axial actions such as waving and walking are generally as low as 2°. More recently, Webber and Rojas (2021) showed the efficacy of using IMUs for human activity recognition without explicitly recovering the orientations (i.e., through gyroscopes). While IMU data could be valuable information for multimodal head and body orientation approaches, existing resources (Alameda-Pineda et al., 2016; Cabrera-Quiros et al., 2018) that focus on social scenes only contain accelerometer data, which is not enough for orientation recovery.

One approach to obtain estimations of head and body orientations is to use the proximity and audio information. Proximity sensors and microphones are already incorporated in the implementation of wearable badges that are common in the social signal processing and affective computing community (i.e., sociometric badges - Olguın and Pentland 2007, OpenBadge - Lederman et al. 2017, etc.). In turn, head and body orientations can be indirectly extracted. Previous work (Alameda-Pineda et al., 2015) used subject ground positions along with speaker/non-speaker correlations and proximity pings to estimate labels of head and body orientations, respectively. Compared to orientation labels from video or IMU, these estimated labels are less reliable since they are derived information from noisy sources. Nonetheless, they can still be explored and it is the focus of this paper.

3. Overview of the approach

Our approach combines 4 kinds of inputs: 1) head and body visual features extracted from head and body image patches, 2) estimated head orientation labels from audio recordings, 3) estimated body orientation labels from infrared proximity sensors, and 4) manually annotated labels of some, but not all, frames. Note that the subject ground positions are assumed to be given for acquiring inputs 2 and 3. The goal of this study is to jointly predict head and body orientations as an 8-class classification problem (dividing 360° into

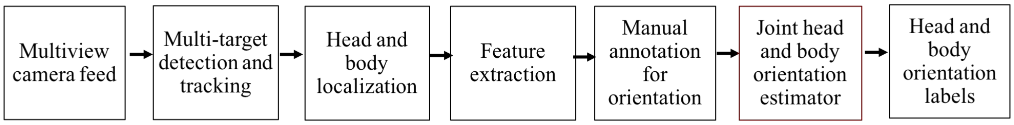

Multiview camera feed
Multi-target detection and tracking
Head and body localization
Feature extraction
Manual annotation for orientation
Joint head and body orientation estimator
Head and body orientation labels

Figure 2: Overall work flow of automatic orientation estimation. The focus of this paper is outlined in red.

eight sectors) using matrix completion in a transductive learning setting. Matrix completion attempts to fill in missing entries in a matrix, which correspond to unobserved orientation labels. It is often solved by iterative optimization. Due to the sparsity and noise in the labels, the underlying challenge is to predict the head and body orientations which are temporally smooth. They also have to be consistent with the manual labels, weak labels (from wearable sensors), and the physical constraints that tend to couple the head and body behavior. For the purpose of this study, we consider multi-person tracking in videos, head and body detection, and appearance-based visual feature extraction as upstream tasks (Figure 2). The core of the proposed model (joint head and body orientation estimator in Figure 2) based on matrix completion is discussed in Section 4, followed by details on experimental conditions in Section 5.

4. Proposed Model

In the supervised learning setting for a linear classifier, the objective is to learn the weight matrix $\boldsymbol{W} \in \mathbb{R}^{c\times(d+1)}$ by minimizing the loss on a training set N_{train} as

$$\arg\min_{\boldsymbol{W}} \sum_{i\in N_{\text{train}}} \text{Loss}\left(\boldsymbol{Y}_i, \boldsymbol{W}\begin{bmatrix}\boldsymbol{X}_i\\ 1\end{bmatrix}\right). \tag{1}$$

$\boldsymbol{W}$ maps the d-dimensional features space $\boldsymbol{X} \in \mathbb{R}^{d\times T}$ to the c-dimensional (number of classes) output space $\boldsymbol{Y} \in \mathbb{R}^{c\times T}$ where T denotes the number of samples in time.

When dealing with noisy features and fuzzy labels, previous research (Bomma and Robertson, 2015; Cabral et al., 2011; Goldberg et al., 2010) have empirically shown the practicality of casting a classification problem into a transductive learning setting such as matrix completion. For our specific task, borrowing from the linear classifier setting, a heterogeneous matrix is built by concatenating the orientation labels $\boldsymbol{Y} \in \mathbb{R}^{c\times T}$, visual features $\boldsymbol{X} \in \mathbb{R}^{d\times T}$, and a row of 1's (to model for bias) as

$$\boldsymbol{J} = \begin{bmatrix}\boldsymbol{Y}\\ \boldsymbol{X}\\ \boldsymbol{1}\end{bmatrix}, \tag{2}$$

where $\boldsymbol{J} \in \mathbb{R}^{(c+d+1)\times T}$.

Note that in (2), $\boldsymbol{Y}$ is a vectorized one hot representation of orientation labels. Dividing 360° into eight sectors means that there are eight possible classes and each orientation belongs to one of the eight classes. For example, an angle θ that is $45^\circ \leq \theta < 90^\circ$would be indicated by the vector $[0, 1, 0, 0, 0, 0, 0, 0]^\top \in \mathbb{R}^{c\times 1}$. Head and body label matrices are

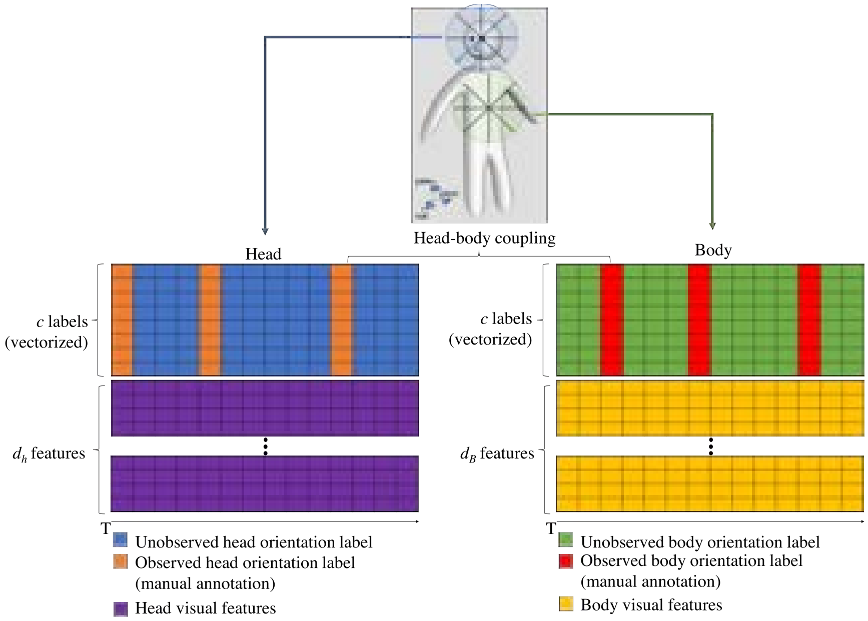

Head
Head-body coupling
Body
c labels (vectorized)
d_h features
c labels (vectorized)
d_B features
T
T
Unobserved head orientation label
Observed head orientation label (manual annotation)
Head visual features
Unobserved body orientation label
Observed body orientation label (manual annotation)
Body visual features

Figure 3: Graphical representation of the feature-label matrix. Head and body orientations are detemined by 2D projections of yaw orientations.

denoted by $\boldsymbol{Y}_h \in \mathbb{R}^{c \times T}$ and $\boldsymbol{Y}_b \in \mathbb{R}^{c \times T}$ respectively. The feature matrices $\boldsymbol{X}_h \in \mathbb{R}^{d_h \times T}$ and $\boldsymbol{X}_b \in \mathbb{R}^{d_b \times T}$ contain the visual features from head and body crops of each person, where d_h and d_b denote the respective feature dimensionality. Following the definition in (2), the visual features and corresponding labels are concatenated into two heterogeneous matrices $\boldsymbol{J}_h = \left[\boldsymbol{Y}_h^\top, \boldsymbol{X}_h^\top, \mathbf{1}^\top\right]^\top$ and $\boldsymbol{J}_b = \left[\boldsymbol{Y}_b^\top, \boldsymbol{X}_b^\top, \mathbf{1}^\top\right]^\top$ for head and body orientation estimation respectively (Figure 3). In addition, a projection matrix $\boldsymbol{P}_h = [\boldsymbol{I}^{cT \times cT}, \mathbf{0}^{cT \times (d_h+1)T}]$ is introduced to extract only the head orientation labels from the heterogeneous matrix $\boldsymbol{J}_h$. In a similar manner, a projection matrix $\boldsymbol{P}_b = [\boldsymbol{I}^{cT \times cT}, \mathbf{0}^{cT \times (d_b+1)T}]$ is defined to extract body orientation labels.

The unobserved orientation labels can either be initialized by information provided by external sources or simply set to zero. In this study, we take the first option. The initial matrices for head and body orientations are denoted by $\boldsymbol{J}_{0,h}$ and $\boldsymbol{J}_{0,b}$ respectively. The label matrix in $\boldsymbol{J}_{0,h}$, denoted by $\boldsymbol{Y}_h$, is further divided into a training set $\mathbf{Y}_{\text{train},h}$ and a test set $\mathbf{Y}_{\text{test},h}$. Similarly, the label matrix in $\boldsymbol{J}_{0,b}$, denoted by $\boldsymbol{Y}_b$, is divided into $\mathbf{Y}_{\text{train},b}$ and $\mathbf{Y}_{\text{test},b}$. Each training set consists of observed labels, while the test set consists of labels to be predicted. We assume that the training and test set samples are interleaved, as shown in Figure 3. We chose this assumption because this could be reflective of real-life scenarios of having observed and unobserved samples intermittently. For the sake of brevity, the subsequent discussion focuses on the head orientation matrix. The body orientation matrix and its corresponding optimization formulation are analogous.

The following discussion outlines the proposed matrix completion method based on the aforementioned setting. We formulate it as an optimization problem, consisting of four components: 1) enforcement of feature-label linear dependency, 2) temporal smoothing,

3) regularization by weak labels, and 4) head-body coupling. Each component applies to completing matrices for estimating head and body orientation respectively. The joint completion of the head and body matrices are further explained in Section 4.5.

4.1. Rank minimization

Following the linear classifier assumption from (2), previous work Goldberg et al. (2010) has shown that the matrix $\boldsymbol{J}_h$ should be low rank. The linear classifier in (1) requires that there is row dependency in (2), hence low rank. The objective is to recover the missing orientation labels such that the rank of the heterogeneous matrix $\boldsymbol{J}_h$ is minimized. Rank minimization is a non-convex problem (Goldberg et al., 2010). However, Candes and Tao (Candes and Tao, 2010) showed that $\text{rank}(\boldsymbol{J}_h)$ can be relaxed to its tightest convex envelope which is the nuclear norm, $\|\boldsymbol{J}_h\|_*$, i.e.

$$\text{rank}(\boldsymbol{J}_h) \approx \|\boldsymbol{J}_h\|_*. \tag{3}$$

In practice, the optimization problem then becomes a minimization of the nuclear norm of $\boldsymbol{J}_h$.

4.2. Temporal smoothing

If samples in the heterogeneous matrix are temporally sorted, one can take advantage of the temporal structure between the columns. Orientation labels are, to an extent, temporally smooth, as head and body poses are not expected to change drastically within a short time period. This can be seen as a column-wise regularization. An interpolated time series of orientation labels $\tilde{\boldsymbol{Y}}_h$ can be generated using an appropriate interpolation scheme to estimate the unobserved orientation labels. In the proposed method, Gaussian process regression (GPR) is chosen as the interpolation scheme. Also known as Kriging, GPR has the same objective as other regression methods, which is to predict the value of a function at some point using a combination of observed values at other points. Rather than curve fitting using a polynomial function for instance, GPR assumes an underlying random process, more specifically a Gaussian process (Bachoc et al., 2017), from which the observed values are sampled. A new posterior distribution is computed based on the assumed (Gaussian process) prior and Gaussian likelihood functions (Williams, 1998). The Gaussian process prior is characterized by a covariance function which measures the similarity between data points; and thus the choice of a suitable covariance function is an essential component in GPR. More details of Gaussian processes and Kriging can be found in Rasmussen and Williams (2005).

Following this procedure, we denote $\boldsymbol{Y}_{\text{GP},h} \in \mathbb{R}^{c \times T}$ as the label matrix where the missing values are imputed by the prediction of GPR. After acquiring the interpolated labels, a new matrix $\boldsymbol{J}_{\text{GP},h}$ is defined as

$$\boldsymbol{J}_{\text{GP},h} = \begin{bmatrix} \boldsymbol{Y}_{\text{GP},h} \\ \boldsymbol{X}_h \\ \boldsymbol{1} \end{bmatrix}. \tag{4}$$

We introduce an additional squared loss term $\|\boldsymbol{P}_h(\boldsymbol{J}_h - \boldsymbol{J}_{\text{GP},h})\|_F^2$ to the optimization problem, where $\|\cdot\|_F^2$ is the Frobenius norm. It is a regularization to ensure that the predicted labels do not deviate drastically from those obtained using temporal interpolation. The projection matrix $\boldsymbol{P}_h$ ensures that the loss is only considered over the orientation labels.

Note that GPR is an example of a regression method that works well in this setting. Alternative regression methods such as Laplacian smoothing (Alameda-Pineda et al., 2015), piece-wise linear interpolation and polynomial regression can also be applied. Our justification for this choice is presented in Section 6.

4.2.1. Gaussian Process Regression Kernels

The basis of GPR is Gaussian Process (GP). A GP is defined to be a random process $f(t)$ for $t \in T$, such that for every finite subset of selected time steps $\{t_1, t_2, ...t_N\}$, $\{f(t_i); i = 1, 2, \ldots, N\}$ is jointly normally distributed. A GP is necessarily defined by its mean function $m(t) = \mathbb{E}[f(t)]$ and its covariance function, also called kernels, $k[t_i, t_j] = \mathbb{E}[(f(t_i) - m(t_i))(f(t_j - m(t_j))]$. While the mean function is often chosen to be zero, the choice of kernels in GP Regression is critical, and is known to affect performance to a great extent. It controls the degree to which data are smoothed when estimating the unknown function (Paciorek and Schervish, 2004). In GP, the kernel represents distance or similarity between two latent variables $f(t_i)$ and $f(t_j)$ given inputs t_i and t_j, $i \neq j$. Intuitively, It describes how output $f(t_j)$ can be affected by output $f(t_i)$. There are many options for these kernel functions. The radial basis function (RBF) kernel is most commonly used and is represented as follows

$$k(t_i, t_j) = \sigma_f^2 e^{-\frac{1}{2}\frac{(t_i - t_j)^2}{\sigma_l^2}}, \tag{5}$$

where σ_l denotes the characteristic length scale that controls the smoothness of the function and σ_f determines the vertical variation .

Matérn kernels are a class of kernels that provide extra flexibility compared to the RBF kernels in controlling the differentiability of the sample functions drawn from the GP distribution. Matérn kernels are of the form

$$k(t_i, t_j) = \sigma_f^2 \frac{2^{1-\nu}}{\Gamma(\nu)} \left(\frac{\sqrt{2\nu} \mid t_i - t_j \mid}{\sigma_l} \right)^{\nu} K_\nu \left(\frac{\sqrt{2\nu} \mid t_i - t_j \mid}{\sigma_l} \right) \tag{6}$$

where ν is the differentiability parameter and K_ν is the modified Bessel function of the second kind. Sample functions drawn from GP with Matérn kernels are $(\lceil \nu \rceil - 1)$ times differentiable, whereas a GP with RBF kernels lead to sample functions that are infinitely differentiable. The parameter ν is usually chosen to be $\frac{3}{2}$ or $\frac{5}{2}$, and (6) can be simplified, respectively, as

$$K_{\frac{3}{2}}(t_i, t_j) = \sigma_f^2 \left(1 + \frac{\sqrt{3}(t_i - t_j)}{\sigma_l} \right) e^{-\frac{\sqrt{3}(t_i - t_j)}{\sigma_l}}, \tag{7}$$

and

$$K_{\frac{5}{2}}(t_i, t_j) = \sigma_f^2 \left(1 + \frac{\sqrt{5}(t_i - t_j)}{\sigma_l} + \frac{5(t_i - t_j)^2}{3\sigma_l^2} \right) e^{-\frac{\sqrt{5}(t_i - t_j)}{\sigma_l}}. \tag{8}$$

The kernels in (7) and (8) lead to once and twice differentiable sample functions in GP. In the subsequent discussion, we refer to (7) and (8) as Matérn 3/2 kernel and Matérn 5/2 kernel, respectively.

The GP kernel function is often chosen based on a qualitative understanding of the underlying data (Vandenberg-Rodes and Shahbaba, 2015). Though RBF kernels are most commonly used, it has been shown that Matérn kernels are more suitable to model physical processes (Stein, 2012; Mertens et al., 2018). Sample functions tend to be less smooth when using Matérn kernels due to finite differentiability, allowing for more realistic capturing of the process. In the context of head and body orientations, it is unlikely that the unknown function would be very highly differentiable.

4.3. Regularization by weak labels

Weak labels estimated from sources such as wearable sensors could be informative though they might be less precise than ground truth (GT) labels. They could still provide additional information that assists in the classification task. We propose a regularization term that incorporates weak labels of head and body orientation. The regularization term can be written as

$$\|\boldsymbol{P}_{w,h}(\boldsymbol{J}_h - \boldsymbol{J}_{w,h})\|_F^2, \tag{9}$$

where $\boldsymbol{P}_{w,h}$ is a projection map that extracts the portions where weak label readings are available. The formulation of $\boldsymbol{J}_{w,h}$ is analogous to (4), where weak labels are treated as approximations of the actual labels. Note that multiple regularization terms of the same form as (9) can be added to the formulation depending on the number of weak labels sources. This highlights the flexibility and modularity of the proposed model in the context of multimodal head and body orientation estimation.

4.4. Head and body coupling

Previous research (Chen et al., 2011; Alameda-Pineda et al., 2015; Varadarajan et al., 2018) has shown that coupling head and body orientation estimation is advantageous for improving accuracy. The proposed formulation also captures the physical constraints between head and body orientations. Since head and body orientations are jointly estimated, this relation fits in nicely as an additional regularization to the optimization problem. It is reasonable to model that head and body orientations cannot be too different at any given time step. Though hinge loss would probably be more appropriate, the relation can also be captured by squared loss, for the ease of analytical derivation and numerical optimization. The regularization term can therefore be written as $\|\boldsymbol{P}_h\boldsymbol{J}_h - \boldsymbol{P}_b\boldsymbol{J}_b\|_F^2$.

4.5. Optimization problem

To summarize, the entire optimization problem, considering all the regularizations and indicating terms associated with both head and body (described in Sections 4.1-4.3), is given by

$$
\begin{aligned}
&\boldsymbol{J}_h^*, \boldsymbol{J}_b^* \\
&= \arg\min_{\boldsymbol{J}_h, \boldsymbol{J}_b} \underbrace{\nu_h \|\boldsymbol{J}_h\|_* + \nu_b \|\boldsymbol{J}_b\|_*}_{\text{matrix low-rankedness}} \\
&\quad + \underbrace{\frac{\lambda_h}{2} \|\boldsymbol{P}_h(\boldsymbol{J}_h - \boldsymbol{J}_{\mathrm{GP},h})\|_F^2 + \frac{\lambda_b}{2} \|\boldsymbol{P}_b(\boldsymbol{J}_b - \boldsymbol{J}_{\mathrm{GP},b})\|_F^2}_{\text{temporal smoothing}} \\
&\quad + \underbrace{\frac{\gamma_h}{2} \|\boldsymbol{P}_{w,h}(\boldsymbol{J}_h - \boldsymbol{J}_{w,h})\|_F^2 + \frac{\gamma_b}{2} \|\boldsymbol{P}_{w,b}(\boldsymbol{J}_b - \boldsymbol{J}_{w,b})\|_F^2}_{\text{weak label regularization}} \\
&\quad + \underbrace{\frac{\mu}{2} \|\boldsymbol{P}_h \boldsymbol{J}_h - \boldsymbol{P}_b \boldsymbol{J}_b\|_F^2,}_{\text{head-body coupling}}
\end{aligned}
\tag{10}
$$

where ν_h, ν_b, λ_h, λ_b, γ_h, γ_b and μ are weights that control the trade-off between the different terms. The equation in (10) can be solved iteratively by an adapted Alternating Direction Method of Multipliers (ADMM) (Boyd et al., 2011; Alameda-Pineda et al., 2015) to jointly solve the minimization problem for the head and body orientation matrices.

Derivation and implementation details are included in Appendix A of the supplementary material. Note that an advantage of the weak labels regularization is that we don't need to study in great detail the quality of the weak labels beforehand. Hyper-parameter optimization will determine the coefficients such that high quality weak labels boost the performance and low quality weak labels get disregarded automatically in squared loss term in (9).

5. Experiments

This section provides a brief introduction of the SALSA dataset (Alameda-Pineda et al., 2016) that was used to obtain the experimental results, and an overview of the experimental protocol. Note that since the premise of our learning problem is transductive and we target a setting with very small number of training samples and labels as well as using multimodal data, we do not compare our method to existing deep learning methods (for head and body orientation estimation) which rely on (re)training on much larger number of labeled data that contain images only and are not multimodal.

5.1. SALSA dataset analysis

5.1.1. Summary

The SALSA dataset is a multimodel dataset that was captured at a social event that consists of a poster presentation session and a mingling event afterwards, involving 18 participants. For this study, we focus on the video recordings, proximity sensor pings, and audio data of the poster presentation session (∼17 minutes). Ground truth labels of head and body orientation of each participant were manually annotated every 3 seconds. Additional details on annotations can be found in Alameda-Pineda et al. (2016). Head and body orientations were extracted from audio and proximity data respectively, independent of the video (Alameda-Pineda et al., 2015). These are treated as weak labels in our context.

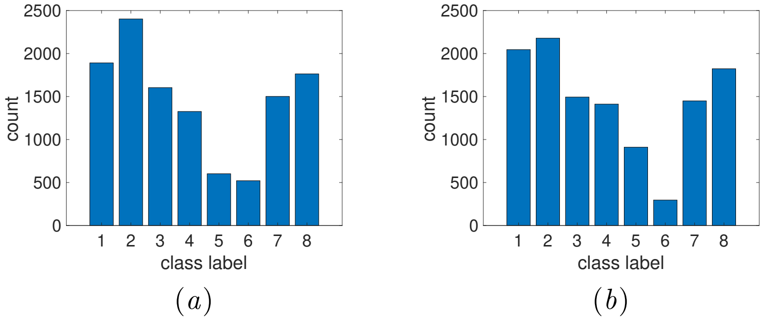


Figure 4: Overall class distribution of head (left) and body (right) GT labels in the SALSA dataset.

The SALSA dataset is a challenging dataset for head and body orientation estimation due to the low resolution of targets, cluttered background, and occlusions. The class distribution of the GT labels is shown in Figure 4. We discretized the GT labels, which are labeled with respect to the ground plane, into 8 angular bins [0,45), [45,90), [90,135), [135,180), [180,225), [225,270), [270,315), and [315,360) in degrees in the room coordinate system; and labeled serially from class 1 to class 8. The majority of GT labels correspond to non-frontal views of the subjects, hence making it difficult to estimate head and body orientations (Varadarajan et al., 2018). Overall, the dataset is relatively balanced except for class 5 and 6. However, person-wise data among the 18 subjects could be heavily imbalanced.

5.1.2. SALSA weak labels analysis

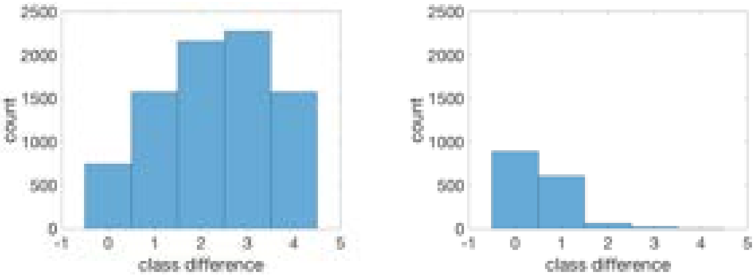
The weak labels estimated for each subject during the poster session of the SALSA dataset are sparse and/or noisy. Head orientation weak labels are extracted by correlating the speaking status between subjects. Body orientation weak labels are extracted based on proximity pings. Both procedures rely on the ground position and relative proximity of the subjects. Weak labels are 28% and 87% sparse for head and body, respectively. Hence, body orientation weak labels are unavailable for most of the poster session. The reason for weak label absence is unclear.

To quantify the quality of the available weak labels, we calculate the difference between the weak labels and GT labels. Since angles are periodic (i.e. repeat every 360°), we take the circular difference δ between the two discretized sets of labels

$$\delta = \min\left(|G_i - W_i|, N - |G_i - W_i|\right), \tag{11}$$

where G_i denotes the i^{th} GT label , W_i the i^{th} weak label, and N the total number of possible classes, which is 8 in the context of this paper. Therefore, the maximal difference does not exceed 4. If the difference is 0, then weak labels match with the GT labels. The distribution of differences in orientation labels are shown via historgrams in Figure 5. As illustrated in the class difference distribution plots, there is generally a considerable discrepancy of class difference of 2 or 3 classes between weak head labels and ground truth. This is expected because microphone data are generally noisy, which can cause errors in estimating speaker and listener status. On the other hand, the class difference in body labels concentrated at

0 is obtained after adding 180° to all weak body labels. This is an artifact that has not been explicitly stated in the original SALSA dataset paper (Alameda-Pineda et al., 2016).



(a) (b)

Figure 5: Distribution of class difference between ground truth and weak labels for head (left) and body (right) orientations.

Poster sessions include moments of high crowd density which compromises the quality of these weak labels, as auditory signals are cross-contaminated and infrared sensors may pick up pings from multiple directions in the vicinity. Previous work (Alameda-Pineda et al., 2015) considered weak labels to be the same quality as GT labels whenever they are available. Also another previous work (Tan et al., 2018) considered head and body orientation estimation as an isolated problem based on only video data. Unlike the aforementioned previous works, this paper exploits the potentially useful information provided by available weak labels. The regularization term in the formulation allows us to circumvent the associated intrinsic noisiness and sparsity (Section 4.3). We also report some investigatory results by simulating labels of different qualities and show how incorporating them via regularization can enhance the model performance. The purpose of this exercise is to provide further insight into future multimodal orientation estimation approaches.

5.2. Experimental setup

We used the Histogram of Gradients (HOG) visual features for head and body crops of each participant from the SALSA dataset poster session, which aligns with the choice in Alameda-Pineda et al. (2015). Similar to the approach proposed by Alameda-Pineda et al. (2015), visual features from the four cameras are concatenated and Principle Component Analysis (PCA) was performed to keep 90% of the variance as dimensionality reduction preprocessing. This results in a 100-dimensional feature vector. Training data are the observed labels and test data are the unobserved labels to be predicted. In a transductive learning setting, since the objective is to predict labels for the unobserved entries only and not generalize to further unseen data, weights are not explicitly learned. Training data and test data partitions are determined by random sampling of columns (over time). Because of this randomness, training and test data are interleaved and we take advantage of this inherent structure in our formulation.

Previously, a person specific training and test scheme, in which a model is trained for every subject, was presented in Tan et al. (2018). A caveat of performance from this type of scheme is that there can be large inter-subject variation. The model trained on one subject may not generalize to other subjects. To investigate the generalizability of the proposed model in this paper, we introduce a person independent training/test protocol. Due to the small subject-wise sample size (18 subjects) of the SALSA dataset, we use a nested leave-one-person-out cross validation (LOPOCV) protocol to conduct the experiments. One subject is left out for each test fold, resulting in 18 folds overall. Within each training fold of 17 subjects, we use a 3-fold cross validation to select the hyperparameters (via Bayesian optimization) in the optimization problem (10). For each subject, the head and body orientation samples are arranged temporally and a random fraction of them are chosen to be training samples. Due to the randomness in this step, we repeat the process of randomly selecting the training samples five times within each of the three folds. We use Bayesian optimization to identify the hyperparameters with the negative of the sum of body and head orientation estimation classification accuracy averaged across the 17 subjects as the objective function.

The model performance on the test subject from each LOPOCV fold is evaluated using the best set of hyperparameters and averaged results from 18 folds are reported. For experimental conditions investigating the influence of the model parameters (Section 6), the model is retrained using the same protocol.

6. Model Analysis

A comprehensive model analysis is conducted considering various possibilities in training schemes, kernel options, and a combination of regularization terms.

6.1. Results

Table 1 reports two sets of baseline results along with results obtained from the proposed model trained using LOPOCV. To obtain the first naive baseline, we simply set the unobserved samples to the value of the mode of the selected samples. The second baseline is the set of person specific results which is reported in Tan et al. (2018). Table 1 shows the averaged-across-subject head and body orientation estimation results for different fractions of manual annotations. There is a notable increase in performance for the proposed model with respect to the two baselines. We also report performance of the proposed model without using the regularization based on the weak labels and observe that including the weak labels has a positive contribution to the performance.

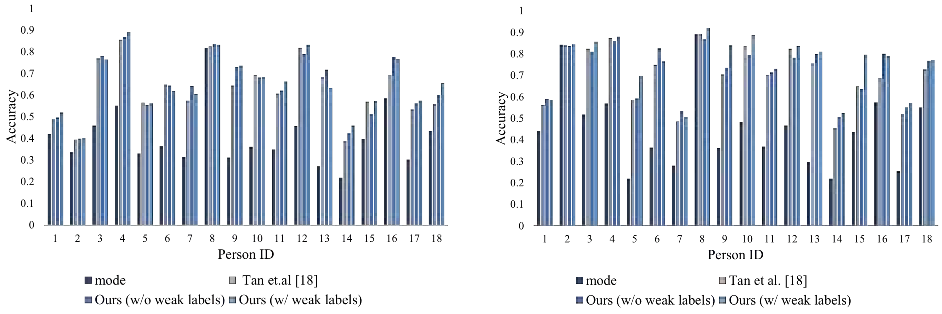
The hyperparameters in the proposed model (10) are $\{\nu_h, \nu_b, \lambda_h, \lambda_b, \gamma_h, \gamma_b, \mu\}$. We arbitrarily set $\nu_b = 1$ as the contribution of the other terms can be considered relative to ν_b. At 5% manual labels, hyperparameter optimization yields $\nu_h = 7.4$, $\lambda_h = 6.4$, $\lambda_b = 5.6$, $\gamma_h = 1.7$, $\gamma_b = 1.3$, and $\mu = 5.2$ averaged across 18 folds of LOPOCV. Comparing ν_h and ν_b, the low rankness of J_h carries more weight than that of J_b in (10). This corroborates the intuition that there is considerable occlusion of subjects' body and less occlusion of subjects' head. We also note that temporal smoothing in both head and body orientations (λ_h and λ_b), and head-body coupling (μ) are important to model performance.

Figure 6 shows a detailed subject-wise comparison at 5% manual labels (i.e., observed samples) fraction. For the majority of the subjects, we notice a consistent improvement with respect to the two baselines. Improvement with respect to the results from Tan et al. (2018) is attributed to the optimization of the GP kernel and weak label regularization

Table 1: Averaged classification accuracy (%) for different fractions (%) of manual annotations. Standard deviation (%) in accuracy performance across all people (in the LOPOCV framework) is shown in the parenthesis. State-of-the-art performance (Alameda-Pineda et al., 2015) at 5% manual annotation is 56.7% and 59.7% for head and body, respectively.

	Fraction	Mode Baseline	Tan et al. Tan et al. (2018)	Ours	
				no weak labels	weak labels
Head	5	40 (13)	63 (13)	64 (13)	65 (13)
	30	41 (13)	68 (13)	72 (13)	74 (13)
	50	41 (13)	70 (13)	77 (9)	76 (12)
	70	41 (13)	71 (13)	77 (11)	77 (12)
Body	5	45 (18)	70 (13)	72 (13)	76 (12)
	30	47 (17)	79 (11)	81 (11)	83 (9)
	50	47 (17)	81 (10)	85 (11)	86 (9)
	70	47 (17)	83 (10)	86 (9)	86 (9)

which were not considered previously. For some subjects such as subject 2 and 8, the mode baseline already performs well, especially for body orientation estimation. This is because orientation variation and diversity are relatively low for these subjects. Larger orientation diversity can lead to lower performance and higher variation across subjects (Tan et al., 2018). On a higher level, this can be related to the personality and role functions of subjects, the dynamics between subjects, and the context of the social scene. For the other manual label fractions, the observations are similar.



(a) (b)

Figure 6: Comparisons of head (left) and body (right) orientation estimation at 5% manual annotation across four setups: mode baseline, Tan et al. (2018), and our formulation without and with regularization by weak labels. The plots are best viewed in color.

6.2. Kernel choice

The choice of the kernel is a critical decision during the modeling process of GPR. Kernel functions encode the underlying behavior of the data such as its periodicity and smoothness. Since we are working with head and body orientation angles, the important feature to take into account is the smoothness. Even though head and body turns could be seen as smooth in general, we hope to capture sudden head and body turns which are more interesting for social scene analysis.

We focus on choosing among the RBF, Matérn 3/2 and Matérn 5/2 kernels. During the hyperparameter optimization, the Matérn 3/2 kernel was found to be the optimal option for all the different fractions of manual annotations listed in Table 1. It further supports with the assumption that head and body orientations are only mildly smooth over time. The RBF kernel assumes that the learned smoothing function is infinitely differentiable which doesn't appear to be as fitting in this particular modeling process. Similarly, the Matérn 5/2 kernel is twice differentiable while the Matérn 3/2 kernel is once differentiable. Further optimization of kernel parameters pertaining to the Matérn 3/2 kernel option was also performed. Signal variance σ_f is a scaling factor that describes the variation of the regressed values to their mean. Characteristic length scale σ_l describes the smoothness of the function. The averaged hyperparameters σ_f and σ_l are 4.6 and 45 respectively.

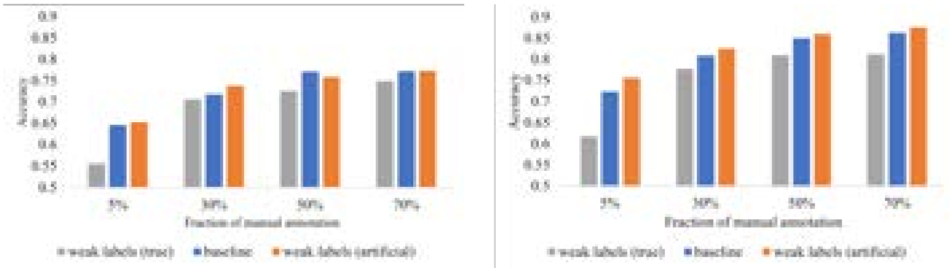
6.3. Regularization by weak labels

In this section, we discuss model performance with two different kinds of weak label inputs for the regularization term in (9). These inputs are used to populate the label portion of $\boldsymbol{J}_{w,h}$ and $\boldsymbol{J}_{w,b}$. First, we use the weak labels provided in the SALSA dataset. Despite the issues with the quality of weak labels as explained in Section 5.1.2, we include the results for instructive purposes. If a weak label is not available at a given timestep, we use the nearest available weak label in time.

The second kind of weak label inputs is artificially generated. We want to investigate how the performance changes with the quality of weak labels. To simulate a set of noisy weak labels, we generate a set of artificial labels by perturbing the GT labels. In practice, we add Gaussian noise with zero mean and standard deviation equal to 15, 30, 60, 90 and 120 degrees. This set of artificial weak labels acts in place of the actual weak labels from SALSA.

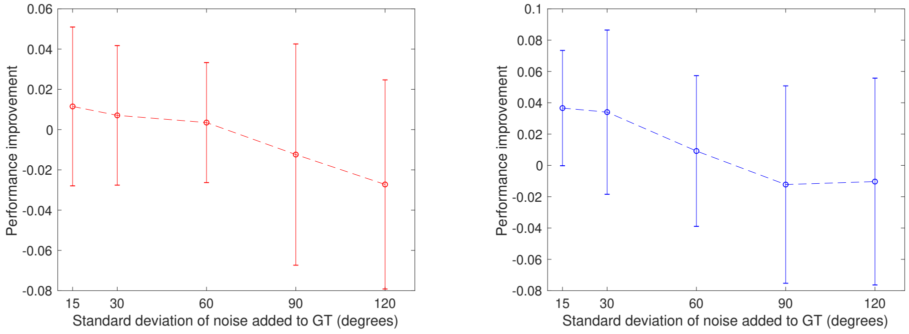
In Figure 7, we report the results obtained with these two types of weak labels. Artificial weak labels have been created with Gaussian noise of standard deviation equal to 30 degrees. The baseline model represents the case when no weak labels are included. We observe that using true weak labels decreases the performance compared to the baseline. This is expected given the poor quality of the actual weak labels. However, with artificial weak labels, there is a notable increase compared to the baseline. This shows that weak labels of decent quality can be exploited, especially when the manual annotation fraction is low. With an increasing number of observed samples, the number of unobserved samples becomes fewer, reducing the dependence on weak labels. As a result, the value of using weak labels diminishes with an increasing number of observed samples. But as we are especially interested in the regime of few observed samples, we highlight the fact that weak labels can indeed boost model performance.

Figure 8 shows the improvement in performance due to noisy weak labels with respect to the baseline model. We set 5% of the data as manual annotations or observed samples. When weak labels become increasingly noisy, the model performance falls below that of the baseline. This demonstrates that weak labels need to be of a reasonable quality to contribute positively to performance and justifies the poor performance when the true SALSA weak labels are included. Furthermore, the improvements in body orientation estimations are more consistent compared to those of the head. Hence, we emphasize that head orientation estimation is a more difficult task, possibly because head orientations vary more than body orientations over short time scales. A different approach such as classification with finer granularities could be promising for better head orientation estimations.



(a) (b)

Figure 7: Performance comparison for head (left) and body (right) orientation estimation without (baseline) and with weak label regularizations. Artificial weak labels have been created using Gaussian noise of standard deviation equal to 30 degrees.

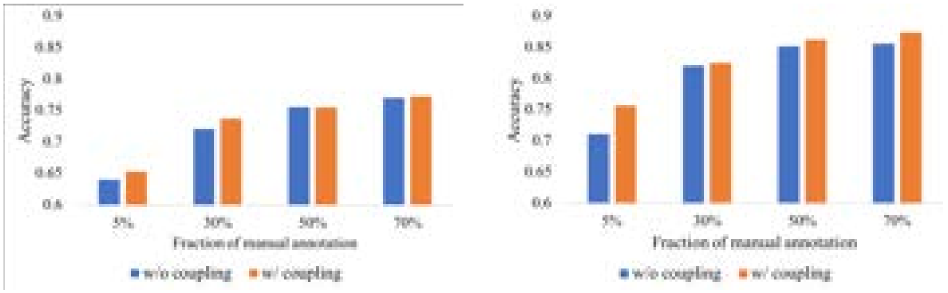


(a) (b)

Figure 8: Improvement in performance of head (left) and body (right) orientation estimation for different magnitudes of noise in artifically generated weak labels. The improvement is reported with respect to the baseline (i.e., no weak labels) in the 5% observed samples setting. The error bars indicate subject wise standard deviation in improvement.

6.4. Contribution of head-body coupling

To study the contribution from head-body coupling regularization term, we remove this from the best model (i.e. with artifical weak labels) and compare the performance difference. Figure 9 shows the extent to which the performance decreases without head-body coupling, which is more prominent when the manual annotation fraction is low. Similar to the observation made for the weak label regularization, when there is more observed



(*a*) (*b*)

Figure 9: Performance comparison for head (left) and body (right) orientation estimation without and with head-body coupling regularization.

samples, GP smoothing becomes advantageous and dominant, making the head-body coupling term less important. However, when the observed sample size is small, the head-body coupling contributes positively to the performance. In particular, the effect is prominent in body orientation estimation where an increase of 4.7% in accuracy is obtained when 5% of the data is manually labeled.

7. Discussion and Conclusion

In this paper, we present a model that utilizes few labeled samples to classify unlabeled samples for head and body orientation estimation in a transductive setting using matrix completion. The formulation of the model combines rank minimization of the joint feature-label matrix, temporal smoothing over labels (based on GPR), weak labels regularization that takes advantage of weak labels from wearable sensors, and head-body coupling to ensure physical restraints of head and body orientation estimates. Since we are especially interested in investigating multimodal orientation estimation, we primarily test our method on the challenging SALSA dataset. SALSA is the largest annotated dataset that contains multiple overlapping video recordings and wearable sensor readings along with ground positions, and head and body orientations of each subject. In Section 5.1, we describe some issues and challenges with working with weak labels acquired from wearable sensors. We do not compare to existing deep learning methods for head and/or body orientation estimation (e.g., Prokudin et al. 2018; Beyer et al. 2015; Raza et al. 2018; Choi et al. 2016) because of the fundamental difference in learning setting and the lack of multimodal comparisons. Future extension of studies based on deep learning approaches could be developed to accomodate multimodal data for this task, upon further ablation studies to verify the efficiency of wearable sensing data.

Notable conclusions from our experimental results are – (*i*) the person independent model achieved by the proposed formulation outperforms the person specific model, which shows promising generalization ability; (*ii*) a more suitable kernel for GPR when modeling head and body orientation series is the Matérn 3/2 kernel, as opposed to the more popular RBF kernel; (*iii*) weak labels of low quality may impair performance but in the case where better quality weak labels are used, model performance is boosted; and (*iv*) head and body coupling indeed improves head and body orientation. The increase in performance due to (*iii*) anf (*iv*) is especially notable in the few manual annotations or observed samples regime.

There are some limitations to this model. The performance would depend on the spacing (availability) of observed samples in order for temporal smoothing to be effective. The method does not apply to independent and isolated unseen samples. It would not perform well if the period of interest is far away in time compared to the observed samples. On the other hand, this provides initial guidelines on selecting which samples to annotate if there are financial constraints. Performance would also depend on the methods applied to the sensor signals as acquiring head and body orientation estimates from wearable sensors is challenging in itself.

Future work entails addressing the aforementioned limitations. On the other hand, given the flexibility of the model, possible topics to explore include but are not limited to matrix completion with missing features, feature representation across different modalities, and joint head and body matrix completion of several subjects, given prior information such as group membership assignments. In the case of a large number of unlabeled samples in a dataset, results from the proposed model would give competitive rough estimates of the actual labels as a data augmentation technique. This is a viable option if obtaining manual labels becomes expensive or impossible. Acquiring results from the model is relatively computationally inexpensive, and we can use them as a springboard for deep neural networks or other models that require a larger number of labeled samples to achieve better head and body orientation estimations.

Acknowledgments

This research was partially funded by the Netherlands Organization for Scientific Research (NWO) under project number 639.022.606. The authors thank Xavier Alameda-Pineda for sharing data and implementation of his previous research (Alameda-Pineda et al., 2015).

References

Hamad Ahmed and Muhammad Tahir. Improving the accuracy of human body orientation estimation with wearable IMU sensors. *IEEE Transactions on instrumentation and measurement*, 66(3):535–542, 2017.

X. Alameda-Pineda, J. Staiano, R. Subramanian, L. Batrinca, E. Ricci, B. Lepri, O. Lanz, and N. Sebe. Salsa: A novel dataset for multimodal group behavior analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 38(8):1707–1720, Aug 2016. ISSN 0162-8828. doi: 10.1109/TPAMI.2015.2496269.

Xavier Alameda-Pineda, Yan Yan, Elisa Ricci, Oswald Lanz, and Nicu Sebe. Analyzing free-standing conversational groups: A multimodal approach. In *Proceedings of the 23rd ACM International Conference on Multimedia*, MM '15, pages 5–14, New York, NY, USA, 2015. ACM. ISBN 978-1-4503-3459-4. doi: 10.1145/2733373.2806238. URL http://doi.acm.org/10.1145/2733373.2806238.

Sileye O Ba and Jean-Marc Odobez. Recognizing visual focus of attention from head pose in natural meetings. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, 39(1):16–33, 2009.

F. Bachoc, F. Gamboa, J. M. Loubes, and N. Venet. A gaussian process regression model for distribution inputs. *IEEE Transactions on Information Theory*, pages 1–1, 2017. ISSN 0018-9448. doi: 10.1109/TIT.2017.2762322.

Valentin Bazarevsky, Ivan Grishchenko, Karthik Raveendran, Tyler Zhu, Fan Zhang, and Matthias Grundmann. Blazepose: On-device real-time body pose tracking. *arXiv preprint arXiv:2006.10204*, 2020.

Lucas Beyer, Alexander Hermans, and Bastian Leibe. Biternion nets: Continuous head pose regression from discrete training labels. In *German Conference on Pattern Recognition*, pages 157–168. Springer, 2015.

S. Bomma and N. M. Robertson. Joint classification of actions with matrix completion. In *2015 IEEE International Conference on Image Processing (ICIP)*, pages 2766–2770, Sept 2015. doi: 10.1109/ICIP.2015.7351306.

Stephen Boyd, Neal Parikh, Eric Chu, Borja Peleato, Jonathan Eckstein, et al. Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends® in Machine learning*, 3(1):1–122, 2011.

Ricardo S. Cabral, Fernando Torre, Joao P. Costeira, and Alexandre Bernardino. Matrix completion for multi-label image classification. In J. Shawe-Taylor, R. S. Zemel, P. L. Bartlett, F. Pereira, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 24*, pages 190–198. Curran Associates, Inc., 2011. URL http://papers.nips.cc/paper/4419-matrix-completion-for-multi-label-image-classification.pdf.

L. Cabrera-Quiros, A. Demetriou, E. Gedik, L. v. d. Meij, and H. Hung. The matchnmingle dataset: a novel multi-sensor resource for the analysis of social interactions and group dynamics in-the-wild during free-standing conversations and speed dates. *IEEE Transactions on Affective Computing*, pages 1–1, 2018. ISSN 1949-3045. doi: 10.1109/TAFFC.2018.2848914.

E. J. Candes and T. Tao. The power of convex relaxation: Near-optimal matrix completion. *IEEE Transactions on Information Theory*, 56(5):2053–2080, May 2010. ISSN 0018-9448. doi: 10.1109/TIT.2010.2044061.

Cristian Canton-Ferrer, Carlos Segura, Josep R Casas, Montse Pardas, and Javier Hernando. Audiovisual head orientation estimation with particle filtering in multisensor scenarios. *EURASIP Journal on Advances in Signal Processing*, 2008:32, 2008.

Zhe Cao, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Realtime multi-person 2d pose estimation using part affinity fields. In *CVPR*, 2017.

Nicolo Carissimi, Paolo Rota, Cigdem Beyan, and Vittorio Murino. Filling the gaps: Predicting missing joints of human poses using denoising autoencoders. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 0–0, 2018.

Cheng Chen, Alexandre Heili, and Jean-Marc Odobez. A joint estimation of head and body orientation cues in surveillance video. In *2011 IEEE International Conference on Computer Vision Workshops (ICCV Workshops)*, pages 860–867. IEEE, 2011.

Jinyoung Choi, Beom-Jin Lee, and Byoung-Tak Zhang. Human body orientation estimation using convolutional neural network. *arXiv preprint arXiv:1609.01984*, 2016.

Vasileios Choutas, Georgios Pavlakos, Timo Bolkart, Dimitrios Tzionas, and Michael J Black. Monocular expressive body regression through body-driven attention. In *European Conference on Computer Vision*, pages 20–40. Springer, 2020.

Gabriele Fanelli, Thibaut Weise, Juergen Gall, and Luc Van Gool. Real time head pose estimation from consumer depth cameras. In *Joint Pattern Recognition Symposium*, pages 101–110. Springer, 2011.

Hao-Shu Fang, Shuqin Xie, Yu-Wing Tai, and Cewu Lu. RMPE: Regional multi-person pose estimation. In *ICCV*, 2017.

Amir Ghodrati, Marco Pedersoli, and Tinne Tuytelaars. Is 2d information enough for viewpoint estimation? *Proceedings BMVC 2014*, pages 1–12, 2014.

Andrew Goldberg, Ben Recht, Junming Xu, Robert Nowak, and Xiaojin Zhu. Transduction with matrix completion: Three birds with one stone. In J. D. Lafferty, C. K. I. Williams, J. Shawe-Taylor, R. S. Zemel, and A. Culotta, editors, *Advances in Neural Information Processing Systems 23*, pages 757–765. Curran Associates, Inc., 2010. URL http://papers.nips.cc/paper/3932-transduction-with-matrix-completion-three-birds-with-one-stone.pdf.

Riza Alp Güler, Natalia Neverova, and Iasonas Kokkinos. Densepose: Dense human pose estimation in the wild. *arXiv*, 2018.

Irtiza Hasan, Francesco Setti, Theodore Tsesmelis, Alessio Del Bue, Fabio Galasso, and Marco Cristani. Mx-lstm: mixing tracklets and vislets to jointly forecast trajectories and head poses. *arXiv preprint arXiv:1805.00652*, 2018.

M. Higger, M. Akcakaya, and D. Erdogmus. A robust fusion algorithm for sensor failure. *IEEE Signal Processing Letters*, 20(8):755–758, Aug 2013. ISSN 1070-9908. doi: 10.1109/LSP.2013.2266254.

Liang Hu, Yujie Tang, Zhipeng Zhou, and Wei Pan. Reinforcement learning for orientation estimation using inertial sensors with performance guarantee. In *2021 IEEE International Conference on Robotics and Automation (ICRA)*, pages 10243–10249. IEEE, 2021.

Weiming Hu, Tieniu Tan, Liang Wang, and S. Maybank. A survey on visual surveillance of object motion and behaviors. *IEEE Transactions on Systems, Man, and Cybernetics,* *Part C (Applications and Reviews)*, 34(3):334–352, Aug 2004. ISSN 1094-6977. doi: 10.1109/TSMCC.2004.829274.

Hung-Hsuan Huang, Naoya Baba, and Yukiko Nakano. Making virtual conversational agent aware of the addressee of users' utterances in multi-user conversation using nonverbal information. In *Proceedings of the 13th international conference on multimodal interfaces*, pages 401–408. ACM, 2011.

Catalin Ionescu, Dragos Papava, Vlad Olaru, and Cristian Sminchisescu. Human3. 6m: Large scale datasets and predictive methods for 3d human sensing in natural environments. *IEEE transactions on pattern analysis and machine intelligence*, 36(7):1325–1339, 2013.

Adam Kendon. *Conducting interaction: Patterns of behavior in focused encounters*, volume 7. CUP Archive, 1990.

Manon Kok and Thomas B Schön. A fast and robust algorithm for orientation estimation using inertial sensors. *IEEE Signal Processing Letters*, 26(11):1673–1677, 2019.

Manon Kok, Jeroen D Hol, and Thomas B Schön. Using inertial sensors for position and orientation estimation. *arXiv preprint arXiv:1704.06053*, 2017.

Sven Kreiss, Lorenzo Bertoni, and Alexandre Alahi. PifPaf: Composite fields for human pose estimation. *arXiv preprint arXiv:1903.06593*, 2019.

Daniel Laidig, Marco Caruso, Andrea Cereatti, and Thomas Seel. Broad—a benchmark for robust inertial orientation estimation. *Data*, 6(7):72, 2021.

Stephen RH Langton, Roger J Watt, and Vicki Bruce. Do the eyes have it? Cues to the direction of social attention. *Trends in cognitive sciences*, 4(2):50–59, 2000.

Oren Lederman, Dan Calacci, Angus MacMullen, Daniel C Fehder, Fiona E Murray, and Alex'Sandy' Pentland. Open badges: A low-cost toolkit for measuring team communication and dynamics. *arXiv preprint arXiv:1710.01842*, 2017.

Donghoon Lee, Ming-Hsuan Yang, and Songhwai Oh. Head and body orientation estimation using convolutional random projection forests. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2017.

Wu Liu, Yongdong Zhang, Sheng Tang, Jinhui Tang, Richang Hong, and Jintao Li. Accurate estimation of human body orientation from rgb-d sensors. *IEEE Transactions on cybernetics*, 43(5):1442–1452, 2013.

Benoît Massé, Silèye O. Ba, and Radu Horaud. Tracking gaze and visual focus of attention of people involved in social interaction. *CoRR*, abs/1703.04727, 2017.

Benoît Massé, Silèye Ba, and Radu Horaud. Tracking gaze and visual focus of attention of people involved in social interaction. *IEEE transactions on pattern analysis and machine intelligence*, 40(11):2711–2724, 2018.

FG Mertens, A Ghosh, and LVE Koopmans. Statistical 21-cm signal separation via Gaussian process regression analysis. *Monthly Notices of the Royal Astronomical Society*, 478(3): 3640–3652, 2018.

Benjamin A Newman, Reuben M Aronson, Siddartha S Srinivasa, Kris Kitani, and Henny Admoni. HARMONIC: A multimodal dataset of assistive human-robot collaboration. *arXiv preprint arXiv:1807.11154*, 2018.

Kaoruko Okuno, Takayoshi Yamashita, Hiroshi Fukui, Shuzo Noridomi, Koji Arata, Yuji Yamauchi, and Hironobu Fujiyoshi. Body posture and face orientation estimation by convolutional network with heterogeneous learning. In *2018 International Workshop on Advanced Image Technology (IWAIT)*, pages 1–4. IEEE, 2018.

Daniel Olguın Olguın and Alex Sandy Pentland. Sociometric badges: State of the art and future applications. In *Doctoral colloquium presented at IEEE 11th International Symposium on Wearable Computers, Boston, MA*, 2007.

Christopher J Paciorek and Mark J Schervish. Nonstationary covariance functions for Gaussian process regression. In *Advances in neural information processing systems*, pages 273–280, 2004.

Sergey Prokudin, Peter Gehler, and Sebastian Nowozin. Deep directional statistics: Pose estimation with uncertainty quantification. *arXiv preprint arXiv:1805.03430*, 2018.

Carl Edward Rasmussen and Christopher K. I. Williams. *Gaussian Processes for Machine Learning (Adaptive Computation and Machine Learning)*. The MIT Press, 2005. ISBN 026218253X.

Mudassar Raza, Zonghai Chen, Saeed-Ur Rehman, Peng Wang, and Peng Bao. Appearance based pedestrians' head pose and body orientation estimation using deep learning. *Neurocomputing*, 272:647–659, 2018.

Yu Rong, Takaaki Shiratori, and Hanbyul Joo. Frankmocap: Fast monocular 3d hand and body motion capture by regression and integration. *arXiv preprint arXiv:2008.08324*, 2020.

István Sárándi, Timm Linder, Kai Oliver Arras, and Bastian Leibe. Metrabs: Metric-scale truncation-robust heatmaps for absolute 3d human pose estimation. *IEEE Transactions on Biometrics, Behavior, and Identity Science*, 3(1):16–30, 2020.

Francesco Setti, Chris Russell, Chiara Bassetti, and Marco Cristani. F-formation detection: Individuating free-standing conversational groups in images. *PloS one*, 10(5):e0123783, 2015.

Fumito Shinmura, Daisuke Deguchi, Ichiro Ide, Hiroshi Murase, and Hironobu Fujiyoshi. Estimation of human orientation using coaxial rgb-depth images. In *VISAPP (2)*, pages 113–120, 2015.

Leonid Sigal and Michael J Black. Measure locally, reason globally: Occlusion-sensitive articulated pose estimation. In *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, volume 2, pages 2041–2048. IEEE, 2006.

Michael L Stein. *Interpolation of spatial data: some theory for kriging*. Springer Science & Business Media, 2012.

Ramanathan Subramanian, Yan Yan, Jacopo Staiano, Oswald Lanz, and Nicu Sebe. On the relationship between head pose, social attention and personality prediction for unstructured and dynamic group interactions. In *Proceedings of the 15th ACM on International Conference on Multimodal Interaction*, ICMI '13, pages 3–10, New York, NY, USA, 2013. ACM. ISBN 978-1-4503-2129-7. doi: 10.1145/2522848.2522862. URL http://doi.acm.org/10.1145/2522848.2522862.

Stephanie Tan, David M.J. Tax, and Hayley Hung. Improving temporal interpolation of head and body pose using Gaussian process regression in a matrix completion setting. In *Proceedings of the Group Interaction Frontiers in Technology*, page 3. ACM, 2018.

Stephanie Tan, David MJ Tax, and Hayley Hung. Multimodal joint head orientation estimation in interacting groups via proxemics and interaction dynamics. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 5(1):1–22, 2021.

Jonathan J Tompson, Arjun Jain, Yann LeCun, and Christoph Bregler. Joint training of a convolutional network and a graphical model for human pose estimation. In Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 27*, pages 1799–1807. Curran Associates, Inc., 2014.

Diego Tosato, Mauro Spera, Marco Cristani, and Vittorio Murino. Characterizing humans on riemannian manifolds. *IEEE transactions on pattern analysis and machine intelligence*, 35(8):1972–1984, 2012.

Alexander Toshev and Christian Szegedy. Deeppose: Human pose estimation via deep neural networks. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2014.

Alexander Vandenberg-Rodes and Babak Shahbaba. Dependent matérn processes for multivariate time series. *arXiv preprint arXiv:1502.03466*, 2015.

Jagannadan Varadarajan, Ramanathan Subramanian, Samuel Rota Bulò, Narendra Ahuja, Oswald Lanz, and Elisa Ricci. Joint estimation of human pose and conversational groups from social scenes. *International Journal of Computer Vision*, 126(2-4):410–429, 2018.

Marton Veges and A Lőrincz. Temporal smoothing for 3d human pose estimation and localization for occluded people. In *International Conference on Neural Information Processing*, pages 557–568. Springer, 2020.

Mitchell Webber and Raul Fernandez Rojas. Human activity recognition with accelerometer and gyroscope: A data fusion approach. *IEEE Sensors Journal*, 21(15):16979–16989, 2021.

Shih-En Wei, Varun Ramakrishna, Takeo Kanade, and Yaser Sheikh. Convolutional pose machines. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

Christopher KI Williams. Prediction with Gaussian processes: From linear regression to linear prediction and beyond. In *Learning in graphical models*, pages 599–621. Springer, 1998.

Chenyan Wu, Yukun Chen, Jiajia Luo, Che-Chun Su, Anuja Dawane, Bikramjot Hanzra, Zhuo Deng, Bilan Liu, James Z Wang, and Cheng-hao Kuo. Mebow: Monocular estimation of body orientation in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3451–3461, 2020.