Improving temporal interpolation of head and body pose using Gaussian process regression in a matrix completion setting

Tan, Stephanie; Tax, David M.J.; Hung, Hayley

DOI
10.1145/3279981.3279982

Publication date
2018

Document Version
Final published version

Published in
Proceedings of the Group Interaction Frontiers in Technology, GIFT 2018

Improving Temporal Interpolation of Head and Body Pose using Gaussian Process Regression in a Matrix Completion Setting

Stephanie Tan, David M.J. Tax, Hayley Hung
Delft University of Technology, Netherlands
{S.Tan-1,D.M.J.Tax,H.Hung}@tudelft.nl

ABSTRACT

This paper presents a model for head and body pose estimation (HBPE) when labelled samples are highly sparse. The current state-of-the-art multimodal approach to HBPE utilizes the matrix completion method in a transductive setting to predict pose labels for unobserved samples. Based on this approach, the proposed method tackles HBPE when manually annotated ground truth labels are temporally sparse. We posit that the current state of the art approach oversimplifies the temporal sparsity assumption by using Laplacian smoothing. Our final solution uses : i) Gaussian process regression in place of Laplacian smoothing, ii) head and body coupling, and iii) nuclear norm minimization in the matrix completion setting. The model is applied to the challenging SALSA dataset for benchmark against the state-of-the-art method. Our presented formulation outperforms the state-of-the-art significantly in this particular setting, e.g. at 5% ground truth labels as training data, head pose accuracy and body pose accuracy is approximately 62% and 70%, respectively. As well as fitting a more flexible model to missing labels in time, we posit that our approach also loosens the head and body coupling constraint, allowing for a more expressive model of the head and body pose typically seen during conversational interaction in groups. This provides a new baseline to improve upon for future integration of multimodal sensor data for the purpose of HBPE.

CCS CONCEPTS

• **Computing methodologies** → **Scene understanding**; • **Theory of computation** → *Semi-supervised learning*;

KEYWORDS

head and body pose estimation, matrix completion

ACM Reference Format:
Stephanie Tan, David M.J. Tax, Hayley Hung. 2018. Improving Temporal Interpolation of Head and Body Pose using Gaussian Process Regression in a Matrix Completion Setting. In *Group Interaction Frontiers in Technology (GIFT'18), October 16, 2018, Boulder, CO, USA.* ACM, New York, NY, USA, Article 4, 8 pages. https://doi.org/10.1145/3279981.3279982

GIFT'18, October 16, 2018, Boulder, CO, USA

ACM ISBN 978-1-4503-6077-7/18/10.
https://doi.org/10.1145/3279981.3279982

(a) (b) (c) (d)

Figure 1: Examples of HBPE challenges from the SALSA dataset [1]. (a) Low resolution (b) Low visibility (c) background clutter (d) occlusion

1 BACKGROUND

Pose estimation has been a popular subject of interest within the computer vision community. While deep learning based state-of-the-art pose estimation methods [17, 30, 31, 33] have achieved remarkable results in articulated pose estimation (i.e. detection and prediction of the location of body parts and joints), pose estimation remains challenging particularly for crowded scenes in the surveillance setting. Hence, it is limited to only head and body pose estimation (HBPE). Despite the seeming simplification of the task, challenges of HBPE in this particular setting [20] include but are not limited to low resolution, low light visibility, background clutter and occlusions (see Figure 1 for example).

HBPE is traditionally a vision-only task. But to tackle these challenges, researchers can leverage on a multi-view camera and multi-sensor scenario [1]. The multi-view camera setting provides multiple perspectives of people in the scene to acquire better HBPE. More interestingly, wearable sensors such as microphones, infrared or bluetooth proximity sensors, etc. have shown an ability to recover HBPE independent from the video modality [22]. Additionally, they can provide more fine-grained information of the human subjects that would not otherwise be available from video only. More specifically, studying small group interactions in crowded space can benefit from data of multiple modalities [15]. In combination with video, these wearable sensors provide a multimodal platform to study detailed and rich information about the human subjects by complementing and enhancing HBPE, which is particularly crucial to the analysis of group and crowd behavior.

Even though it would be ideal to combine multiple modalities, wearable sensors such as microphones and infrared proximity sensors which have previously been used in form of sociometers, are less reliable and noisier compared to surveillance video footage for the purpose of HBPE. Another problem is that malfunctions of wearable sensors are more difficult to notice compared to those of video cameras, especially during real-time data collection where

camera functionality can be more easily confirmed visually. Due to the difficulties of working with wearable sensors, the resulting data can be either partial or entirely missing [19]. Given that working with a patchwork of multimodal data can be hard, in this paper, we exploit them as part of an initialization step and focus on the problem of interpolating between sparse labels.

The setting of this study is that: i) there is a relatively small number of head and body pose samples ($\sim 10^2-10^3$) for each subject, ii) we want to predict pose labels for unobserved samples only using a very small number ($\sim 5\%$) of sparsely distributed ground truth labels, and iii) we want to take advantage of the temporal structure within the pose label data. A deep learning based method that takes into account this setting will perform sub-optimally due to small number of training samples, and also require extensive computational power and hyperparameter tuning. On the other hand, a matrix completion based transductive learning method which is more explainable and less computationally expensive, addresses the problem setting adequately. Inspired and building upon previous work by Alameda-Pineda et al. [2], the contributions of this study are: i) an enhanced temporal smoothing scheme based on Gaussian process regression for label propagation, and ii) a more interpretable person-wise pose label prediction implementation in the transductive setting using matrix completion.

2 RELATED WORK

Head pose estimation (HPE) and body pose estimation (BPE) have primarily been studied by the computer vision community [27]. While impressive results can be achieved using end-to-end deep learning architectures when using datasets capturing frontal face [25] or the full body [11], HPE and BPE remain to be challenging when dealing with wide angle surveillance, with low resolution, heavy occlusions of targets, and cluttered backgrounds. The problem is often reduced to an 8-class classification problem (dividing 360° into eight sectors), though formulating HBPE as a regression problem [32] or being able to reduce coarseness in estimations can provide more meaningful information for higher level social tasks, such as predicting direction of social attention [24] and personality traits [29]. Pioneering work [3, 12, 28] in HPE and BPE saw first successes of these tasks 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 a person's direction of movement, one line of work focuses on the joint estimation of head and body pose to achieve improved results [12]. Overall, there is more previous work on HPE compared to BPE in the surveillance and crowded setting. In this particular setting, human heads can be more easily seen and therefore head poses are easier to predict. Typically, HPE under this setting already provides rich enough information for high level tasks [3]. On the other hand, human bodies can be occluded when the camera view is from an elevated angle, which makes body orientations more difficult to predict without side information. In contrast, HBPE in other contexts such as AR/VR video gaming, sports, etc. where full frontal body poses are available, is much more well-studied and can be represented by a considerable number of work (e.g. [11], [21]). Hasan et al. [18] have recently proposed a noteworthy deep learning method based on Long Short Term Memory (LSTM) neural networks to jointly forecast trajectories and head poses. This work points to the possibility of utilizing LSTM models to predict head and body pose sequences, which is a more challenging but also more descriptive task, compared to solving HPE and BPE in a classification setting using Convolutional Neural Networks (CNN) [23]. For the specific setting of this paper where annotations are sparsely available, the choice of using deep learning models may not be suitable.

Figure 2: Overall work flow of this study. The focus of this study is highlighted in yellow

In this paper, we propose to use matrix completion for HBPE, which was first proposed by Alameda-Pineda et al. [2]. This approach combines head and body visual features, inferred head orientation labels from audio recordings, body orientation labels from infrared proximity sensors, and manually annotated labels of some but not all frames. To reduce the manual effort of annotating the head and body poses, labels were only created every 3 seconds. Alameda-Pineda et al. poses the estimation of head and body orientations as a matrix completion problem where the visual features and labels from either wearable sensors or manual annotations are concatenated into a heterogeneous matrix, for head and body respectively. Due to sparsity and noise in the data extracted from the wearable sensors, the underlying challenge is to construct a matrix that is temporally smooth; and that is consistent with the manual annotations, the observed wearable sensor readings, and the physical constraints that tend to couple the head and body behaviour together.

3 OUR APPROACH

The scope of the study is to jointly predict head and body pose labels as an 8-class classification problem (dividing 360° into eight sectors) in a matrix completion transductive learning setting. Before performing HBPE, upstream processes such as multi-person detection and tracking in videos, head and body localization, and appearance-based visual feature extraction are carried out as outlined in Figure 2 [2, 12]. The construction of a matrix consisting of visual features and manually annotated labels is illustrated in Figure 3. Head pose features and labels are arranged into one such matrix, and similarly for body pose features and labels. Head and pose labels of each participant (independent of other participants) are estimated by completing their head and body matrices jointly. The technical core of constructing such matrices for HBPE and jointly completing the head and body matrices using our formulation is discussed in Section 4, followed by details on experimental conditions pertaining to the upstream processes (see blue modules in Figure 2) in Section 5.

4 METHODOLOGY

In the supervised learning setting for a linear classifier, the objective is to learn the weight matrix $\boldsymbol{W} \in \mathbb{R}^{c\times(d+1)}$, which maps the d-dimensional features space $\boldsymbol{X} \in \mathbb{R}^{d\times T}$ to the c-dimensional (number

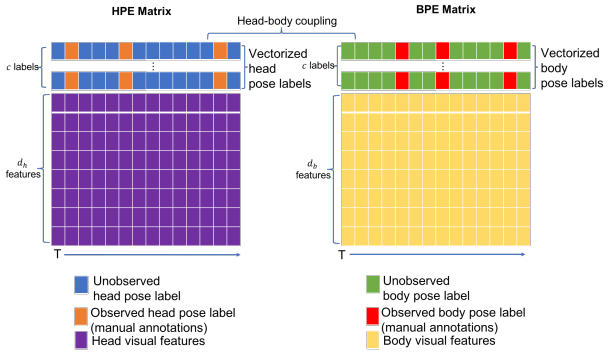

Figure 3: Graphical representation of the feature-label matrix

of classes) output space $\boldsymbol{Y} \in \mathbb{R}^{c \times T}$ where T denotes the number of samples in time, by minimizing the loss on a training set N_{train} as

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

When dealing with noisy features and fuzzy labels, previous research by Bomma and Robertson [5], Cabral et al. [7] and Goldberg et al. [16] have empirically shown the practicality of casting a classification problem into a transductive learning setting such as matrix completion. To that purpose, borrowing from the linear classifier setting, a heterogeneous matrix can be built by concatenating the pose 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 $\boldsymbol{J} \in \mathbb{R}^{(c+d+1) \times T}$.

In the HBPE setting, the duration that we are interested in predicting the pose estimations for is indicated by T and this is represented by arranging samples column-wise for temporal consistency. Note that in (2), $\boldsymbol{Y}$ is a vectorized one hot representation of pose labels. Dividing 360° into eight sectors means that there are eight possible classes and each pose belongs to one of the eight classes. For example, a pose angle between 45° and 90° would be indicated by the vector $[0, 1, 0, 0, 0, 0, 0, 0]^\top \in \mathbb{R}^{c \times 1}$. The head and body label matrices are 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 heterogeneous matrices are $\boldsymbol{J}_h = \left[\boldsymbol{Y}_h^\top, \boldsymbol{X}_h^\top, \boldsymbol{1}^\top\right]^\top$ and $\boldsymbol{J}_b = \left[\boldsymbol{Y}_b^\top, \boldsymbol{X}_b^\top, \boldsymbol{1}^\top\right]^\top$ for head pose and body pose estimation respectively. In addition, a projection matrix $\boldsymbol{P}_h = [\boldsymbol{I}^{c \times c}, \boldsymbol{0}^{c \times (d_h+1)}]$ is introduced to extract only the head pose labels from the heterogeneous matrix $\boldsymbol{J}_h$. In a similar manner, a projection matrix $\boldsymbol{P}_b = [\boldsymbol{I}^{c \times c}, \boldsymbol{0}^{c \times (d_b+1)}]$ is defined to extract body pose labels.

Matrix completion is a formulation that attempts to fill in missing entries in a matrix, which in our context correspond to unobserved pose labels. Matrix completion is often solved by iterative optimization. For the purpose of the iterative scheme, the unobserved pose labels can either be initialized by side information provided by external sources, or simply set to zeros. In this study, we take the first option by initializing the unobserved samples by sensor data. The initial matrices for head and body poses 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 $\boldsymbol{Y}_{\text{train},h}$ and a test set $\boldsymbol{Y}_{\text{test},h}$. Similarly, the label matrix in $\boldsymbol{J}_{0,b}$, denoted by $\boldsymbol{Y}_b$, is divided into $\boldsymbol{Y}_{\text{train},b}$ and $\boldsymbol{Y}_{\text{test},b}$. Each training set consists of observed labels, while the test set consists of unobserved labels. The training set and test set samples are interleaved, as shown in Figure 3. In this study, training set labels are sampled from manual annotations and test set labels are initialized by sensor data, in the hope of achieving faster convergence. For the sake of brevity, the subsequent discussion will be explained for the head pose matrix. The body pose matrix and its corresponding optimization formulation are analogous to those of the head pose matrix.

The following discussion outlines the proposed matrix completion method based on the aforementioned setting. The proposed method consists of three components: i) nuclear norm minimization, ii) temporal smoothing, and iii) head-body coupling.

4.1 Nuclear norm minimization

Following the linear classifier assumption from (2), Goldberg et al. [16] have shown that the matrix $\boldsymbol{J}$ should be low rank. More concretely, the objective is to recover the missing pose labels such that the rank of the heterogeneous matrix $\boldsymbol{J}$ is minimized. Rank minimization is a non-convex problem [16]. However, Candes and Tao [10] showed that $\text{rank}(\boldsymbol{J})$ can be relaxed to its tightest convex envelope which is the nuclear norm, $\|\boldsymbol{J}\|_*$, i.e.

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

The optimization problem then becomes a minimization of the nuclear norm of $\boldsymbol{J}$.

4.2 Temporal smoothing

If samples in the heterogeneous matrix are temporally sorted, one can take advantage of the temporal structure between the columns. Pose labels are to a certain extent, temporally smooth, as poses are not expected to change drastically within a short time period. This can be seen as a column-wise regularization. Using the training set $\boldsymbol{Y}_{\text{train}}$, an interpolated time series of pose labels $\tilde{\boldsymbol{Y}}$ can be generated using an appropriate interpolation scheme to estimate the unobserved pose labels entirely based on temporal consideration. 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 a 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 distribution [4], from which the observed values are sampled. A new posterior distribution is computed based on the assumed (Gaussian process) prior and Gaussian likelihood functions [34]. 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. For

the purpose of this study, the covariance function is chosen to be the popular Radial-Basis Function (RBF) kernel. More details of Gaussian processes and Kriging can be found in [26].

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

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

A squared loss term $\|\boldsymbol{P}(\boldsymbol{J} - \boldsymbol{J}_{\text{GP}})\|_F^2$ is introduced into the nuclear norm minimization problem for regularization to ensure that the predicted labels do not deviate drastically from the labels obtained as a result of temporal interpolation. The projection matrix $\boldsymbol{P}$ ensures that the loss is only considered over the pose labels.

Note that GPR is an example of a regression method that works well in this setting. Alternative regression methods such Laplacian smoothing [2], piece-wise linear interpolation and polynomial regression can also be applied. Our justification of this choice follows in the discussion section in Section 7.

4.3 Head and body coupling

So far the formulation details the manipulation of HPE and BPE matrices separately. In this section we jointly consider the two matrices as they are related. Previous research by Alameda-Pineda et al. [2], Chen et al. [13] and Varadarajan et al. [32] has shown that coupling HPE and BPE is advantageous for improving accuracy. The proposed formulation also captures the physical constraints between head and body poses. Since head and body pose 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 poses 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 $\|\boldsymbol{P}_h\boldsymbol{J}_h - \boldsymbol{P}_b\boldsymbol{J}_b\|_F^2$.

4.4 Optimization problem

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

$$\begin{aligned} \boldsymbol{J}_{h*}, \boldsymbol{J}_{b*} = \arg\min_{\boldsymbol{J}_h, \boldsymbol{J}_b} \; & \nu_h \|\boldsymbol{J}_h\|_* + \nu_b \|\boldsymbol{J}_b\|_* \\ & + \frac{\lambda_h}{2}\|\boldsymbol{P}_h(\boldsymbol{J}_h - \boldsymbol{J}_{\text{GP},h})\|_F^2 + \frac{\lambda_b}{2}\|\boldsymbol{P}_b(\boldsymbol{J}_b - \boldsymbol{J}_{\text{GP},b})\|_F^2 \\ & + \frac{\mu}{2}\|\boldsymbol{P}_h\boldsymbol{J}_h - \boldsymbol{P}_b\boldsymbol{J}_b\|_F^2, \end{aligned} \tag{5}$$

where ν_h, ν_b, λ_h, λ_b, and μ are weights that control the trade-off between the different terms. The equation in (5) can be solved iteratively by an adapted Alternating Direction Method of Multipliers (ADMM) proposed by Boyd et al. [6] and Alameda-Pineda et al. [2] to jointly solve the minimization problem for the head and body pose matrices. We adopt the aforementioned algorithm that starts with the construction of the augmented Lagrangian, similar to the classical ADMM [14]. The augmented Lagrangian of the optimization problem in (5) is given by

$$\begin{aligned} \mathcal{L} = \; & \nu_h \|\boldsymbol{J}_h\|_* + \nu_b \|\boldsymbol{J}_b\|_* \\ & + \frac{\lambda_h}{2}\|\boldsymbol{P}_h(\boldsymbol{K}_h - \boldsymbol{J}_{\text{GP},h})\|_F^2 + \frac{\lambda_b}{2}\|\boldsymbol{P}_b(\boldsymbol{K}_b - \boldsymbol{J}_{\text{GP},b})\|_F^2 \\ & + \frac{\mu}{2}\|\boldsymbol{P}_h\boldsymbol{K}_h - \boldsymbol{P}_b\boldsymbol{K}_b\|_F^2 \\ & + \frac{\phi_h}{2}\|\boldsymbol{K}_h - \boldsymbol{J}_h\|_F^2 + \frac{\phi_b}{2}\|\boldsymbol{K}_b - \boldsymbol{J}_b\|_F^2 \\ & + \langle \boldsymbol{M}_h, \boldsymbol{J}_h - \boldsymbol{K}_h \rangle + \langle \boldsymbol{M}_b, \boldsymbol{J}_b - \boldsymbol{K}_b \rangle, \end{aligned} \tag{6}$$

where $\boldsymbol{K}_h$ and $\boldsymbol{K}_b$ are auxiliary variables that allow the decoupling of the optimization of $\boldsymbol{J}_h$ and $\boldsymbol{J}_b$; and $\boldsymbol{M}_h$ and $\boldsymbol{M}_b$ are Lagrange Multiplier matrices. The inner product of the two terms is denoted by $\langle \cdot, \cdot \rangle$. The update rules are similar to those of the ADMM with scaled dual variables [6]. In this context, the update rules at the k-th iteration are given by

$$\begin{aligned} (\boldsymbol{J}_h^{k+1}, \boldsymbol{J}_b^{k+1}) = \arg\min_{\boldsymbol{J}_h^k, \boldsymbol{J}_b^k} \; & \nu_h \|\boldsymbol{J}_h^k\|_* + \nu_b \|\boldsymbol{J}_b^k\|_* \\ & + \frac{\phi_h}{2}\|\boldsymbol{K}_h^k - \boldsymbol{J}_h^k\|_F^2 + \frac{\phi_b}{2}\|\boldsymbol{K}_b^k - \boldsymbol{J}_b^k\|_F^2 \\ & + \langle \boldsymbol{M}_h^k, \boldsymbol{J}_h^k - \boldsymbol{K}_h^k \rangle + \langle \boldsymbol{M}_b^k, \boldsymbol{J}_b^k - \boldsymbol{K}_b^k \rangle \end{aligned} \tag{7}$$

$$\begin{aligned} (\boldsymbol{K}_h^{k+1}, \boldsymbol{K}_b^{k+1}) = \arg\min_{\boldsymbol{K}_h^k, \boldsymbol{K}_b^k} \; & \frac{\lambda_h}{2}\|\boldsymbol{P}_h(\boldsymbol{K}_h^k - \boldsymbol{J}_{\text{GP},h})\|_F^2 \\ & + \frac{\lambda_b}{2}\|\boldsymbol{P}_b(\boldsymbol{K}_b^k - \boldsymbol{J}_{\text{GP},b})\|_F^2 \\ & + \frac{\mu}{2}\|\boldsymbol{P}_h\boldsymbol{K}_h^k - \boldsymbol{P}_b\boldsymbol{K}_b^k\|_F^2 \\ & + \frac{\phi_h}{2}\|\boldsymbol{K}_h^k - \boldsymbol{J}_h^{k+1}\|_F^2 \\ & + \frac{\phi_b}{2}\|\boldsymbol{K}_b^k - \boldsymbol{J}_b^{k+1}\|_F^2 \\ & + \langle \boldsymbol{M}_h^k, \boldsymbol{J}_h^{k+1} - \boldsymbol{K}_h^k \rangle + \langle \boldsymbol{M}_b^k, \boldsymbol{J}_b^{k+1} - \boldsymbol{K}_b^k \rangle \end{aligned} \tag{8}$$

$$\boldsymbol{M}_h^{k+1} = \boldsymbol{M}_h^k + \phi_h(\boldsymbol{J}_h^{k+1} - \boldsymbol{K}_h^{k+1}) \tag{9}$$

$$\boldsymbol{M}_b^{k+1} = \boldsymbol{M}_b^k + \phi_b(\boldsymbol{J}_b^{k+1} - \boldsymbol{K}_b^{k+1}) \tag{10}$$

More derivation and implementation details can be found in Appendix A.

5 EXPERIMENTAL SETUP

This section provides a brief introduction of the SALSA dataset that was used to obtain the experimental results, and an overview of the experimental conditions.

5.1 SALSA Dataset

The SALSA dataset is captured at a social event that consists of a poster presentation session and a mingling event afterwards, involving 18 participants. It is a multimodal dataset that includes video recordings from a multi-view surveillance camera (4 cameras) network, binary proximity sensor data acquired from sociometric badges worn by the participants, and audio recordings of each participant acquired by a microphone embedded in the sociometric badges. For this study, we only focus on the video recordings of the poster presentation session. Ground truth labels of head and body pose of each participant were manually annotated every 3 seconds.

There are in total 645 ground truth annotations for each head and body pose of each participant. The authors of [2] also inferred head pose from microphone data and body pose from infrared proximity sensor data, independent from the video modality. These are considered as "soft" labels and further details of their extraction can be found in [2] and are provided as part of the dataset.

5.2 Experimental Conditions

We used the provided Histogram of Gradients (HOG) visual features for head and body crops of each participant from the SALSA dataset poster session [2]. Similar to Alameda-Pineda et al. [2]'s approach, visual features from the four cameras are concatenated and PCA was performed to keep 90% of the variance. 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, it is conventional to have both the training data and test data available during training. Because 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 defined by a random projection mask to simulate random sampling over labels. Because of this randomness, training and test data are interleaved and we take advantage of this inherent structure in our formulation. Note that because of the same reason, all our experiments are conducted 20 times and results are averaged to mediate the random projection masks causing overestimation or underestimation of prediction accuracies. Additionally, the sample diversity (i.e. class distribution) is different among participants. Hence, a randomly created projection mask is rejected if it results in low sample diversity in the training set. The hyper-parameters in (6) are optimized using Bayesian optimization with 5-fold cross validation.

5.3 Implementation Details

Similar to the authors of [2], we assume visual features from each participant are available at any given time step. Unlike in previous approach [2] where the inferred "soft" labels are used as part of the training set, our experiments only use samples that were manually annotated to construct the training set. It is unclear if the experiments reported by Alameda-Pineda et al. [2] used additional unlabeled samples along with the manual annotations and "soft" labels in their model during training.

Since we were not able to clarify ambiguities in the description of the experimental setup in the former formulation [2], we made the following decisions regarding the experimental setting. In this study, we construct the training and test sets from only samples that are manually annotated in the SALSA dataset. Since the quality of the "soft" labels was not quantitatively assessed by Alameda-Pineda et al. [2], in our case, it also makes sense for us to avoid training using "soft" labels so we can more clearly see the effect of our proposed approach independently of the influence of training with weak labels. In our experiments, although "soft" labels are not considered as part of the observed samples, they are only used as initializations of unobserved samples in order to reach faster convergence. Note that columns of the matrix which are initially populated with soft labels are subject to immediate changes after being fed as inputs to the optimization problem.

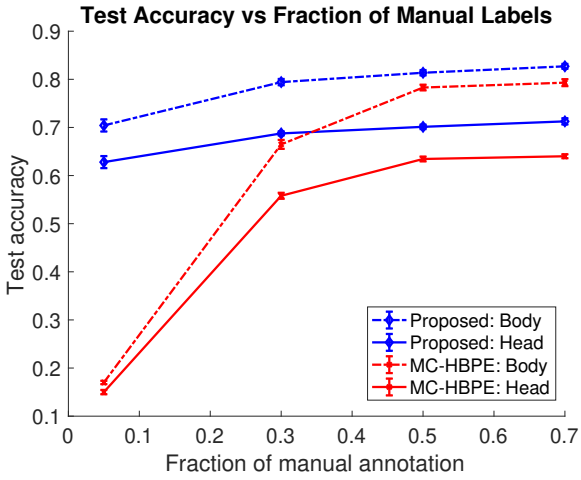

Figure 4: Test accuracy of HPE and BPE using MC-HBPE [2] and the proposed method. Error bars indicate the standard deviations of results for each fraction of manual annotation.

6 RESULTS

The heterogeneous matrix for head and body are initialized with the same fraction of ground truth labels as training data, though their respective random projection masks are different. Figure 4 shows the test accuracy, which is the prediction accuracy over unobserved labels, against different fractions of manual annotation used for training. The proposed method is compared against the state-of-the-art matrix completion based HBPE method by Alameda-Pineda et al. [2]. As shown in Figure 4, the proposed method is drastically superior compared to the state-of-the-art matrix completion by Alameda-Pineda et al. [2], especially at very low fraction of manual annotations.

The difference in performance of both methods is accredited to a simple numerical phenomenon. One of the major differences between the proposed method and the method by Alameda-Pineda et al. [2] is the temporal smoothing scheme. In the latter, the authors employed Laplacian smoothing to ensure temporal consistency over the pose estimates. While it is a reasonable choice for smoothing based on local information, GPR in contrast provides smoothing by exploiting a more global context based on only a few data points. By fitting sparse data points in the functional space, GPR is known to better recover nonlinear patterns and longer timescale trends compared to polynomial interpolation, and especially Laplacian smoothing. As a result, it provides a good accuracy even when only 5% of the manual labels are available as training data. Additionally, person-wise HBPE results for all 18 participants at 5% manual annotation using the two methods is reported in Table 1.

During social events and in free-standing conversation groups, we expect head pose to change more than body poses. Hence, it is reasonable to conclude that head poses are harder to predict compared to body poses; and it is reflected in the observation that test accuracies for head pose estimates are lower than test accuracies for body pose estimates from both the methods. This can be further analyzed by computing the entropy to illustrate the distribution

Manual Annotation: 5%	MC-HBPE[2]		Proposed		Labels diversity (Entropy)	
	HPE mean (std)	BPE mean (std)	HPE mean (std)	BPE mean (std)	Head	Body
Person 1 [119]	0 (0)	0 (0)	0.49 (3.2e-2)	0.56 (4.4e-2)	1.18	1.13
Person 2 [132]	0.06 (1.9e-3)	0 (0)	0.39 (1.0e-2)	0.84 (1.3e-2)	1.30	0.51
Person 3 [140]	0.63 (3.1e-2)	0.68 (3.3e-2)	0.77 (2.0e-2)	0.82 (2.7e-2)	1.48	1.28
Person 4 [169]	0.02 (1.7e-3)	0.01 (1.2e-3)	0.86 (2.7e-2)	0.87 (2.2e-2)	1.19	1.10
Person 5 [177]	0.13 (2.3e-3)	0.25 (2.4e-2)	0.57 (5.3e-2)	0.58 (6.4e-2)	1.78	1.74
Person 6 [180]	0.44 (2.1e-2)	0.39 (1.7e-2)	0.65 (3.5e-2)	0.75 (4.6e-2)	1.67	1.53
Person 7 [216]	0.17 (6.5e-2)	0.19 (2.7e-2)	0.57 (5.4e-2)	0.48 (7.2e-2)	1.72	1.84
Person 8 [238]	0.01 (9.1e-4)	2.5e-4 (5.9e-4)	0.82 (1.1e-2)	0.89 (1.9e-2)	0.63	0.41
Person 9 [241]	0.43 (4.1e-3)	0.57 (3.7e-3)	0.64 (6.4e-2)	0.70 (6.9e-2)	1.53	1.55
Person 10 [261]	0.09 (2.7e-3)	0.23 (2.5e-3)	0.69 (2.9e-2)	0.85 (3.5e-2)	1.37	1.20
Person 11 [262]	0.13 (2.0e-3)	0.01 (2.2e-3)	0.61 (3.5e-2)	0.70 (4.5e-2)	1.52	1.47
Person 12 [267]	0.13 (1.1e-3)	0.03 (1.1e-3)	0.82 (2.2e-2)	0.83 (2.0e-2)	1.01	0.96
Person 13 [286]	0 (0)	0 (0)	0.68 (2.9e-2)	0.76 (3.0e-2)	1.61	1.56
Person 14 [307]	0.09 (1.7e-2)	0.12 (3.7e-2)	0.39 (4.2e-2)	0.46 (6.8e-2)	1.82	1.74
Person 15 [313]	0 (0)	0 (0)	0.57 (7.0e-2)	0.65 (6.8e-2)	1.16	1.06
Person 16 [350]	0.03 (1.9e-3)	0.03 (2.9e-2)	0.69 (6.3e-2)	0.69 (6.1e-2)	1.22	1.22
Person 17 [351]	0.26 (4.1e-2)	0.29 (3.3e-2)	0.53 (3.5e-2)	0.52 (3.9e-2)	1.69	1.69
Person 18 [353]	0.13 (2.2e-2)	0.20 (1.1e-3)	0.56 (5.5e-2)	0.73 (6.0e-2)	1.39	1.11

* [·] indicates the person ID encoding provided in the SALSA dataset.

Table 1: Person-Wise HBPE results using MC-HBPE [2] and the proposed method. Difficulty of HBPE for each person is captured quantitatively in labels diversity measured by labels entropy.

of the ground truth labels used in this study. The equation for calculating entropy is given by

$$H = -\sum_{i=1}^{c} P_i \log P_i, \tag{11}$$

where H is the information entropy measure of a set of samples and P_i is the proportion of ground truth labels in the i^{th} class. For an unbiased 8 class label distribution (i.e. uniform distribution), the maximum entropy value is approximately 2.08. The entropy of head pose labels averaged over all participants is 1.4. The entropy of body pose labels averaged over all participants is 1.28. Therefore, head pose diversity is slightly higher than that of the body pose for 16/18 subjects in this dataset, which partially justifies the reasoning that head pose labels are more difficult to accurately predict than body pose labels within this dataset. However, the GPR-based proposed method still manages to achieve significantly higher test accuracies for head pose estimates compared to the method by Alameda-Pineda et al. [2].

It is worth noting that in this study, we sample training data from manual labels, whereas in the experimental setup by Alameda-Pineda et al. [2], "soft" labels acquired from wearable devices are also used as part of training data. Experiments were also conducted where the "soft" labels provided in the dataset are included as part of the training data. However, no desirable results can be obtained. As a reference, using the same approach as that of [2] at 50% training data partition with 5% manual annotations and 95% "soft" labels, we obtained 14% and 16% for HPE and BPE respectively, compared to the reported 57% and 60% [2].

7 DISCUSSION

In our proposed method, GPR performs fitting over the head and body pose estimates separately, which loosens the head and body coupling constraint to a certain extent. Though there is still point to point coupling between head pose and body pose at each time step, the head poses and body poses are each separately governed by their own trend which should be less sensitive to noise from the other. Coupling that is too tight may artificially enforce head and body pose to be the same which may not reflect the reality when it comes to small group interactions. This implicit benefit from recovering nonlinearities independently should provide rich information to study human behavior in groups.

Since wearable sensors are known to provide noisy information, not all "soft" labels can be seen to have the same quality as ground truth labels. It would be ideal to add high quality "soft" labels to training data and if they are as high quality as manual labels, they can further benefit and improve HBPE in a multimodal setting, as opposed to a single video modality. However, this prior knowledge would need to be obtained beforehand. Because the proposed formulation gives robust performance with small number of manual annotations without the use of any "soft" labels, it provides a good baseline and ground for comparison for further investigation of the quality of labels derived from wearable sensors.

While the highlight of this formulation is to predict the classification of unobserved labels based on a very small number of observed labels, the model does not extend to predicting further unseen data since the weights are not explicitly recovered. When an observed label becomes available to be included, the full model

needs to be run again. One of the computational bottlenecks is Gaussian process regression, which has $O(n^3)$ time complexity that makes it infeasible to scale up for large quantities of data. Another computational bottleneck is the singular value decomposition (SVD) in solving the optimization problem using ADMM (see Appendix A).

8 CONCLUSION

This work focuses on estimating head and body poses in crowded social scene scenario using Gaussian process regression and head and body coupling as a regularization term in a matrix completion setting. The model's premise is to predict head and body pose labels as an 8-class classification problem in a transductive learning setting. The model is able to predict a relatively large percentage of pose labels in large continuous time segment (average 20 samples gap length, approximately 1 minute in real time) and implicitly recover the nonlinearity within the data using only a small fraction of samples as training data. The proposed model has shown to be effective on the challenging SALSA dataset and achieved desirable results of 62% accuracy on head pose estimation and 70% accuracy on body pose estimation using only 5% of the samples as training data, showing superior performance over the state-of-the-art.

Future work on improving HBPE includes integrating wearable sensor data as regularization terms towards a truly multimodal approach. Rather than using appearance based HOG features, visual features could also be extracted using a CNN pre-trained on large image databases and fine-tuned on the SALSA dataset. Additionally, it would be interesting to assess the performance of the proposed method on different, but equally challenging datasets, such as the MatchNMingle dataset [8]. Further analysis of HBPE performance with respect to participants' role in the social scenarios in question and their pose diversity may lend deeper insights to fine-grained head and body movements in group interactions.

9 ACKNOWLEDGMENT

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 [2].

REFERENCES

[1] X. Alameda-Pineda, J. Staiano, R. Subramanian, L. Batrinca, E. Ricci, B. Lepri, O. Lanz, and N. Sebe. 2016. SALSA: A Novel Dataset for Multimodal Group Behavior Analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 38, 8 (Aug 2016), 1707–1720. https://doi.org/10.1109/TPAMI.2015.2496269

[2] Xavier Alameda-Pineda, Yan Yan, Elisa Ricci, Oswald Lanz, and Nicu Sebe. 2015. Analyzing Free-standing Conversational Groups: A Multimodal Approach. In *Proceedings of the 23rd ACM International Conference on Multimedia (MM '15)*. ACM, New York, NY, USA, 5–14. https://doi.org/10.1145/2733373.2806238

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

[4] F. Bachoc, F. Gamboa, J. M. Loubes, and N. Venet. 2017. A Gaussian Process Regression Model for Distribution Inputs. *IEEE Transactions on Information Theory* (2017), 1–1. https://doi.org/10.1109/TIT.2017.2762322

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

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

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

[8] L. Cabrera-Quiros, A. Demetriou, E. Gedik, L. v. d. Meij, and H. Hung. 2018. 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* (2018), 1–1. https://doi.org/10.1109/TAFFC.2018.2848914

[9] J. Cai, E. Candès, and Z. Shen. 2010. A Singular Value Thresholding Algorithm for Matrix Completion. *SIAM Journal on Optimization* 20, 4 (2010), 1956–1982. https://doi.org/10.1137/080738970 arXiv:https://doi.org/10.1137/080738970

[10] E. J. Candes and T. Tao. 2010. The Power of Convex Relaxation: Near-Optimal Matrix Completion. *IEEE Transactions on Information Theory* 56, 5 (May 2010), 2053–2080. https://doi.org/10.1109/TIT.2010.2044061

[11] Zhe Cao, Tomas Simon, Shih-En Wei, and Yaser Sheikh. 2017. Realtime Multi-Person 2D Pose Estimation using Part Affinity Fields. In *CVPR*.

[12] Cheng Chen, Alexandre Heili, and Jean-Marc Odobez. 2011. Combined estimation of location and body pose in surveillance video. In *Advanced Video and Signal-Based Surveillance (AVSS), 2011 8th IEEE International Conference on*. IEEE, 5–10.

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

[14] Jonathan Eckstein and W Yao. 2012. Augmented Lagrangian and alternating direction methods for convex optimization: A tutorial and some illustrative computational results. *RUTCOR Research Reports* 32 (2012), 3.

[15] Daniel Gatica-Perez. 2009. Automatic Nonverbal Analysis of Social Interaction in Small Groups. *Image Vision Comput.* 27, 12 (Nov. 2009), 1775–1787. https://doi.org/10.1016/j.imavis.2009.01.004

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

[17] Riza Alp Güler, Natalia Neverova, and Iasonas Kokkinos. 2018. DensePose: Dense Human Pose Estimation In The Wild. *arXiv* (2018).

[18] Irtiza Hasan, Francesco Setti, Theodore Tsesmelis, Alessio Del Bue, Fabio Galasso, and Marco Cristani. 2018. MX-LSTM: mixing tracklets and vislets to jointly forecast trajectories and head poses. *arXiv preprint arXiv:1805.00652* (2018).

[19] M. Higger, M. Akcakaya, and D. Erdogmus. 2013. A Robust Fusion Algorithm for Sensor Failure. *IEEE Signal Processing Letters* 20, 8 (Aug 2013), 755–758. https://doi.org/10.1109/LSP.2013.2266254

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

[21] Eldar Insafutdinov, Leonid Pishchulin, Bjoern Andres, Mykhaylo Andriluka, and Bernt Schiele. 2016. Deepercut: A deeper, stronger, and faster multi-person pose estimation model. In *European Conference on Computer Vision*. Springer, 34–50.

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

[23] Yang Lu, Shujuan Yi, Nan Hou, Jingfu Zhu, and Tiemin Ma. 2016. Deep neural networks for head pose classification. In *Intelligent Control and Automation (WCICA), 2016 12th World Congress on*. IEEE, 2787–2790.

[24] Benoît Massé, Silèye O. Ba, and Radu Horaud. 2017. Tracking Gaze and Visual Focus of Attention of People Involved in Social Interaction. *CoRR* abs/1703.04727 (2017).

[25] Erik Murphy-Chutorian and Mohan Manubhai Trivedi. 2009. Head pose estimation in computer vision: A survey. *IEEE transactions on pattern analysis and machine intelligence* 31, 4 (2009), 607–626.

[26] Carl Edward Rasmussen and Christopher K. I. Williams. 2005. *Gaussian Processes for Machine Learning (Adaptive Computation and Machine Learning)*. The MIT Press.

[27] Leonid Sigal. 2014. *Human Pose Estimation*. Springer US, Boston, MA, 362–370. https://doi.org/10.1007/978-0-387-31439-6_584

[28] Leonid Sigal and Michael J Black. 2006. Measure locally, reason globally: Occlusion-sensitive articulated pose estimation. In *Computer Vision and Pattern Recognition, 2006 IEEE Computer Society Conference on*, Vol. 2. IEEE, 2041–2048.

[29] Ramanathan Subramanian, Yan Yan, Jacopo Staiano, Oswald Lanz, and Nicu Sebe. 2013. 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)*. ACM, New York, NY, USA, 3–10. https://doi.org/10.1145/2522848.2522862

[30] Jonathan J Tompson, Arjun Jain, Yann LeCun, and Christoph Bregler. 2014. Joint Training of a Convolutional Network and a Graphical Model for Human Pose Estimation. In *Advances in Neural Information Processing Systems 27*, Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger (Eds.). Curran Associates, Inc., 1799–1807.
[31] Alexander Toshev and Christian Szegedy. 2014. DeepPose: Human Pose Estimation via Deep Neural Networks. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.
[32] Jagannadan Varadarajan, Ramanathan Subramanian, Samuel Rota Bulò, Narendra Ahuja, Oswald Lanz, and Elisa Ricci. 2018. Joint Estimation of Human Pose and Conversational Groups from Social Scenes. *International Journal of Computer Vision* 126, 2-4 (2018), 410–429.
[33] Shih-En Wei, Varun Ramakrishna, Takeo Kanade, and Yaser Sheikh. 2016. Convolutional Pose Machines. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.
[34] Christopher KI Williams. 1998. Prediction with Gaussian processes: From linear regression to linear prediction and beyond. In *Learning in graphical models*. Springer, 599–621.

A DERIVATIONS OF ADMM

To separate head and body expressions, at k^{th} iteration, the optimization problem (7) can be split into

$$\boldsymbol{J}_h^{k+1} = \arg\min_{\boldsymbol{J}_h^k} \nu_h \|\boldsymbol{J}_h^k\|_* + \frac{\phi_h}{2}\|\boldsymbol{K}_h^k - \boldsymbol{J}_h^k\|_F^2 + \langle \boldsymbol{M}_h^k, \boldsymbol{J}_h^k - \boldsymbol{K}_h^k\rangle \tag{12}$$

and

$$\boldsymbol{J}_b^{k+1} = \arg\min_{\boldsymbol{J}_b^k} \nu_b \|\boldsymbol{J}_b^k\|_* + \frac{\phi_b}{2}\|\boldsymbol{K}_b^k - \boldsymbol{J}_b^k\|_F^2 + \langle \boldsymbol{M}_b^k, \boldsymbol{J}_b^k - \boldsymbol{K}_b^k\rangle. \tag{13}$$

Simplifying and manipulating (12), we obtain

$$\begin{aligned}\boldsymbol{J}_h^{k+1} = &\arg\min_{\boldsymbol{J}_h^k} \nu_h \|\boldsymbol{J}_h^k\|_* \\ &+ \frac{\phi_h}{2}[\langle \boldsymbol{K}_h^k, \boldsymbol{K}_h^k\rangle - 2\langle \boldsymbol{K}_h^k, \boldsymbol{J}_h^k\rangle + \langle \boldsymbol{J}_h^k \boldsymbol{J}_h^k\rangle] \\ &+ \langle \boldsymbol{M}_h^k, \boldsymbol{J}_h^k\rangle - \langle \boldsymbol{M}_h^k, \boldsymbol{K}_h^k\rangle \\ &+ \frac{1}{2\phi_h}\langle \boldsymbol{M}_h^k, \boldsymbol{M}_h^k\rangle - \frac{1}{2\phi_h}\langle \boldsymbol{M}_h^k, \boldsymbol{M}_h^k\rangle.\end{aligned} \tag{14}$$

Equation (14) can be arranged as

$$\boldsymbol{J}_h^{k+1} = \arg\min_{J_h} \frac{\nu_h}{\phi_h}\|\boldsymbol{J}_h\|_* + \frac{1}{2}\left\|\frac{1}{\phi_h}\boldsymbol{M}_h^k + \boldsymbol{J}_h - \boldsymbol{K}_h^k\right\|_F^2 - \frac{1}{2\phi_h}\langle \boldsymbol{M}_h^k, \boldsymbol{M}_h^k\rangle. \tag{15}$$

The last term in Equation (15) results in a scalar constant and does not affect the nature of optimization. The solution to minimization problem (15) was derived by Cai et al. [9] and Alameda-Pineda et al. [2], and is given by

$$\boldsymbol{J}_h^{k+1} = \boldsymbol{U}_h S_{\frac{\nu_h}{\phi_h}}(\boldsymbol{D}_h)\boldsymbol{V}_h^\top, \tag{16}$$

where the $\boldsymbol{U}_h$, $\boldsymbol{D}_h$, and $\boldsymbol{V}_h$ are obtained from singular value decomposition (SVD) of matrix $\boldsymbol{K}_h^k - \frac{1}{\phi_h}\boldsymbol{M}_h^k$

$$[\boldsymbol{U}_h, \boldsymbol{D}_h, \boldsymbol{V}_h] = \text{SVD}\left(\boldsymbol{K}_h^k - \frac{1}{\phi_h}\boldsymbol{M}_h^k\right) \tag{17}$$

and where the shrinkage operator is given by

$$S_\lambda(x) = \max(x - \lambda, 0) \tag{18}$$

and is applied element-wise to the diagonal matrix of singular values $\boldsymbol{D}_h$. The derivations can be similarly extended for body pose estimation matrix and the solution is given by

$$\boldsymbol{J}_b^{k+1} = \boldsymbol{U}_b S_{\frac{\nu_b}{\phi_b}}(\boldsymbol{D}_b)\boldsymbol{V}_b^\top. \tag{19}$$

For the second step in the optimization problem (8), we define the row-vectorization form of the matrices $\boldsymbol{K}_h$ and $\boldsymbol{K}_b$ as $\boldsymbol{k}_h = \text{vec}(\boldsymbol{K}_h)$ and $\boldsymbol{k}_b = \text{vec}(\boldsymbol{K}_b)$ respectively. The row vectorization notation extends to other matrices in (8) similarly. Derivatives of the objective function in (8) with respect to $\boldsymbol{k}_h$ and $\boldsymbol{k}_b$ are given by

$$\frac{\partial \mathcal{L}}{\partial \boldsymbol{k}_h} = \lambda_h \boldsymbol{P}_h(\boldsymbol{k}_h - \boldsymbol{j}_{GP,h}) + \mu \boldsymbol{P}_h^\top(\boldsymbol{P}_h\boldsymbol{k}_h - \boldsymbol{P}_b\boldsymbol{k}_b) + \phi_h(\boldsymbol{k}_h - \boldsymbol{j}_h^{k+1}) - \boldsymbol{m}_h^k, \tag{20}$$

and

$$\frac{\partial \mathcal{L}}{\partial \boldsymbol{k}_b} = \lambda_b \boldsymbol{P}_b(\boldsymbol{k}_b - \boldsymbol{j}_{GP,b}) + \mu \boldsymbol{P}_b^\top(\boldsymbol{P}_b\boldsymbol{k}_b - \boldsymbol{P}_h\boldsymbol{k}_h) + \phi_b(\boldsymbol{k}_b - \boldsymbol{j}_b^{k+1}) - \boldsymbol{m}_b^k. \tag{21}$$

Equating this derivative to 0 results in a system of linear equations for $\boldsymbol{k}_h^{k+1}$ and $\boldsymbol{k}_b^{k+1}$ given by

$$(\lambda_h \boldsymbol{P}_h + \mu \boldsymbol{P}_h^\top \boldsymbol{P}_h + \phi_h)\boldsymbol{k}_h^{k+1} = \lambda_h \boldsymbol{P}_h \boldsymbol{j}_{GP,h} + \mu \boldsymbol{P}_h^\top \boldsymbol{P}_b \boldsymbol{k}_b + \phi_h \boldsymbol{j}_h^{k+1} + \boldsymbol{m}_h^k \tag{22}$$

and

$$(\lambda_b \boldsymbol{P}_b + \mu \boldsymbol{P}_b^\top \boldsymbol{P}_b + \phi_b)\boldsymbol{k}_b^{k+1} = \lambda_b \boldsymbol{P}_b \boldsymbol{j}_{GP,b} + \mu \boldsymbol{P}_b^\top \boldsymbol{P}_h \boldsymbol{k}_h + \phi_b \boldsymbol{j}_b^{k+1} + \boldsymbol{m}_b^k. \tag{23}$$

Hence, these two equations can be easily solved using standard solvers based on LU decomposition or iterative solvers such as conjugate gradient method to yield the minimizers $\boldsymbol{k}_h^{k+1}$ and $\boldsymbol{k}_b^{k+1}$. We can reshape the solved row vectors $\boldsymbol{k}_h^{k+1}$ and $\boldsymbol{k}_b^{k+1}$ back to matrix forms denoted by $\boldsymbol{K}_h^{k+1}$ and $\boldsymbol{K}_b^{k+1}$. Additionally, the system of linear equations (22) and (23) can be further simplified to give analytical solutions. For the sake of brevity, the reader is referred to the derivation by Alameda-Pineda et al. [2] in their supplementary material.