

Simultaneous Feature and Body-Part Learning for Real-Time Robot Awareness of Human Behaviors

Fei Han¹, Xue Yang¹, Christopher Reardon², Yu Zhang³, and Hao Zhang¹

Abstract—Robot awareness of human actions is an essential research problem in robotics with many important real-world applications, including human-robot collaboration and teaming. Over the past few years, depth sensors have become a standard device widely used by intelligent robots for 3D perception, which can also offer human skeletal data in 3D space. Several methods based on skeletal data were designed to enable robot awareness of human actions with satisfactory accuracy. However, previous methods treated all body parts and features equally important, without the capability to identify discriminative body parts and features. In this paper, we propose a novel simultaneous *Feature And Body-part Learning (FABL)* approach that simultaneously identifies discriminative body parts and features, and efficiently integrates all available information together to enable *real-time* robot awareness of human behaviors. We formulate FABL as a regression-like optimization problem with structured sparsity-inducing norms to model interrelationships of body parts and features. We also develop an optimization algorithm to solve the formulated problem, which possesses a theoretical guarantee to find the optimal solution. To evaluate FABL, three experiments were performed using public benchmark datasets, including the MSR Action3D and CAD-60 datasets, as well as a Baxter robot in practical assistive living applications. Experimental results show that our FABL approach obtains a high recognition accuracy with a processing speed of the order-of-magnitude of 10^4 Hz, which makes FABL a promising method to enable *real-time* robot awareness of human behaviors in practical robotics applications.

I. INTRODUCTION

In a wide variety of human-centered robotics applications, including human-robot teaming, human-robot collaboration, and robot-assisted living, robot awareness of human actions (or behaviors) is essential for intelligent robots to understand humans, make situationally appropriate decisions, and interact with and assist people. However, robot awareness of human behaviors in real-world environments is a challenging problem caused by significant variations of human motion, diversity of human appearance, and vision difficulties, including illumination variations and occlusion. When implemented on robots, additional challenges are encountered, such as uncertainty in movement and dynamic backgrounds; Most importantly, the requirement of real-time performance demands timely robot planning and decision making.

¹Fei Han, Xue Yang, and Hao Zhang are with the Department of Computer Science, Colorado School of Mines, Golden, CO 80401, USA. fhan@mines.edu, edyxueyx@gmail.com, hzhang@mines.edu.

²Christopher Reardon is with the US Army Research Laboratory, Adelphi, MD 20783, USA. christopher.m.reardon3.civ@mail.mil.

³Yu Zhang is with the Department of Computer Science and Engineering, Arizona State University, Tempe, AZ 85281, USA. yzhan442@asu.edu.

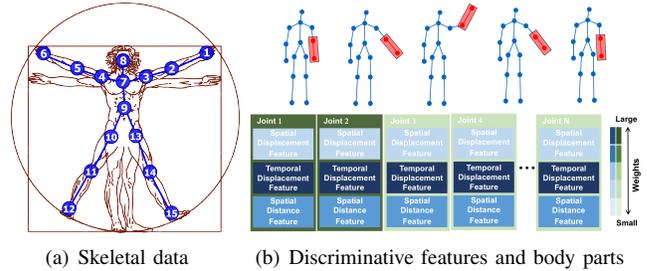


Fig. 1. A motivating example of the FABL approach, which simultaneously learns discriminative skeleton joints and multimodal heterogeneous features to enable real-time robot awareness of human behaviors.

Although human action understanding has been researched in robotics and computer vision communities, most previous techniques are based on local spatio-temporal visual features [1], [2], which are generally incapable of dealing with the challenges introduced by robotics applications (e.g., real-time performance). With the emergence of affordable structured-light or time-of-flight depth sensing technologies, color-depth cameras have generally become a standard 3D visual sensing device for modern indoor robots. The skeletal data of humans acquired from such sensors, as shown in Fig. 1(a), provides the possibility to achieve real-time robot awareness of human behaviors [3], which also provides benefits in comparison to local features, including the invariance to viewpoint, human body scale and motion speed [4], [5].

Because of these advantages, skeleton-based action understanding methods have attracted increasing attention, and many skeletal features and representations have been implemented during the last few years, see [4] and references therein, i.e. joint rotation matrix [6], BIPOD [5], etc. However, most existing methods apply only one type of skeletal feature [7], [6], while others simply concatenate several types of skeletal features together into a single bigger vector to encode human actions [8], [9]. The problem of autonomously learning the importance of skeletal features and optimally integrating the multimodal features (different human activity representations extracted from skeletal data) together has not yet been well addressed for real-time robot awareness of human behaviors. Recently, methods based on *body parts (represented as joints in skeletal data)* instead of using complete skeleton data were studied to improve action recognition accuracy [5], [10], [11]. To remove irrelevant joints for specific behaviors, these methods use a subset of or select skeletal joints. Although these methods obtained promising accuracy, the selection is manual based upon fixed

criteria and is not robust to various scenarios. Furthermore, the question of how to integrate multimodal skeletal features into body-part methods has not been well answered.

In this paper, we introduce a novel *Feature And Body-part Learning* (FABL) method to enable real-time robot awareness of human behaviors, through learning discriminative skeletal features and body parts simultaneously in the same optimization framework. For learning the importance of body parts, our approach is inspired by the insight that typically a subset of body parts are more discriminative to recognize an action. For example, as demonstrated in Fig. 1(b), only the waving arm and hand are important for the action of “hand waving.” Our FABL method is able to select discriminative body parts automatically for different behaviors. Simultaneously, FABL learns the importance of heterogeneous skeletal features, and integrates multimodal features to build a more discriminative representation to enable robot awareness of human behaviors. Classification is seamlessly integrated in the FABL approach (i.e., no external classifier is required), which further increases processing efficiency, resulting in high-speed performance that is suitable for applications with real-time requirements.

The contributions of this paper are twofold:

- We propose a novel formulation and the FABL approach to perform simultaneous learning of discriminative body parts and skeletal features for real-time robot awareness of human behaviors.
- We develop a new optimization algorithm to efficiently solve the formulated robot learning problem, which has a theoretical guarantee to converge to the global optimal solution.

We make the code that implements our FABL approach available at: <http://hcr.mines.edu/code/FABL.html>.

The remainder of this paper is structured as follows. Related work is described in Section II. Then, our FABL approach is detailed in Sections III and IV. Experimental results are presented in Section V. After discussing several attributes of the proposed FABL method in Section VI, we conclude this paper in Section VII.

II. RELATED WORK

In this section, we conduct a review of techniques to understand human actions using skeletal data, including both complete skeletal data and partial body parts.

A. Behavior Understanding Based on Skeletal Data

Methods using 3D skeletal data to identify human actions attracted increasing attention after the release of the affordable structured-light 3D sensing technology [4]. A widely applied representation for human action understanding is based on skeletal joint displacements. Chen and Koskela [12] implemented a feature extraction method based on pairwise relative position of skeletal joints with normalization, and actions were classified by multiple extreme learning machines. Wei *et al.* [13] implemented a hierarchical graph to represent spatio-temporal joint positions and displacements, where the differences in skeletal joint positions between two successive

frames were defined as features. Besides joint displacements, many methods based on joint orientations were also implemented. Sung *et al.* [6] computed the orientation matrix of each joint with respect to the camera, then transformed the matrix to obtain this joint orientation with respect to the human torso, showing their representation was invariant to the sensor’s location. Another popular category of skeleton-based methods directly use raw joint position information for human action understanding. Wei *et al.* [14] developed wavelet features to represent a sequence of 3D skeletal joints, and a concurrent action detection model to understand human behaviors.

Most of the previous skeleton-based methods utilized only one category of skeleton-based features. Several recent studies indicate that recognition accuracy can be improved by combining multiple skeletal features together. A feature construction approach was introduced in [7] that concatenates static posture, movement, and offset values into a single bigger feature vector, and utilizes a naive Bayes classifier to perform multi-class action classification. Yu *et al.* [15] used three categories of skeletal features, including pairwise joint distance, spatial joint coordinate, and temporal variation of joint locations, to construct a mixed representation. A similar skeleton-based representation was implemented by [16], incorporating pairwise joint distances and temporal joint location changes together. However, most previous techniques simply concatenated different categories of features without considering the importance of each skeletal feature category. The research problem of how to autonomously learn and fuse heterogeneous skeletal features for real-time robot awareness of human actions has not yet been well studied.

The proposed FABL approach addresses this problem by integrating heterogeneous multimodal skeletal features through learning the importance of each feature category, along with learning discriminative body parts, to accurately interpret human actions.

B. Representation Based on Body Parts

Skeletal human representations based on body part models have been widely studied in the past few years. Because these mid-level body part models can partially take into account the physical structure of human body, they can yield improved discrimination power to represent humans [5].

Wang *et al.* [17] implemented a method that decomposed a body model into five parts, including left/right arms/legs and the torso, each consisting of a set of joints, to represent human behaviors in space and time dimensions. A spatial-temporal And-Or graph model was implemented in [18] to represent humans at three levels including poses, spatiotemporal-parts, and parts. The hierarchical human body structure captures the geometric and appearance variation of humans at each frame. A deep neural network was introduced in [19] to create a body part model and the correlation of body parts was investigated, which can automatically obtain mid-level features that were more descriptive than low-level features extracted from individual human skeleton

joints. Several methods were also proposed to select more descriptive human body joints [2], [10], [11], [20], [21], [22], [23].

Bio-inspired body part models are also commonly applied to extract mid-level features for skeleton-based representation construction, which are typically based on body kinematics or human anatomy. Chaudhry *et al.* [24] implemented bio-inspired mid-level features to represent human activities based on 3D skeleton data, by leveraging the findings in the research area of static shape encoding in the primate cortex's neural pathway. By showing different 3D shapes to primates and measuring their neural responses, the primates' internal shape representation was estimated, which was then used to extract body parts to create skeleton-based representations. Zhang and Parker [5] proposed a new bio-inspired predictive orientation decomposition representation, which was inspired by the biological research in human anatomy. This approach decomposed a body model into five body parts, and projected 3D human skeleton trajectories onto three anatomical planes. Through estimating future skeleton trajectories, this method is able to predict future human motions.

Despite the promising results obtained by the methods based on body parts, which mutually partition the body model into several body parts or select a set of skeletal joints according to predefined criteria, previous techniques did not model the discrimination difference of human joints but simply include or exclude certain joints. In this paper, we introduce a new approach to automatically learn discriminative skeletal joints without predefined manual selection criteria.

III. THE FABL APPROACH

In this section, we describe our FABL method that simultaneously learns discriminative skeletal features and body parts to enable real-time robot awareness of human behaviors.

Notation. In this paper, we denote matrices using boldface capital letters, and vectors using boldface lowercase letters. We represent the ℓ_1 -norm of a vector $\mathbf{v} \in \mathbb{R}^n$ using $\|\mathbf{v}\|_1 = \sum_{i=1}^n |v_i|$, and the ℓ_2 -norm of \mathbf{v} as $\|\mathbf{v}\|_2 = \sqrt{\mathbf{v}^\top \mathbf{v}}$. Given a matrix $\mathbf{M} = \{m_{ij}\} \in \mathbb{R}^{m \times n}$, we refer to its i -th row as \mathbf{m}^i and the j -th column as \mathbf{m}_j . We denote the Frobenius norm of the matrix \mathbf{M} as $\|\mathbf{M}\|_F = \sqrt{\sum_{i=1}^m \sum_{j=1}^n m_{ij}^2}$.

A. Problem Formulation

Given a collection of n data instances, the skeletal matrix is denoted as $\mathbf{X} = [\mathbf{x}_1, \dots, \mathbf{x}_n] \in \mathbb{R}^{d \times n}$, where $\mathbf{x}_i \in \mathbb{R}^d$ is the vector of all skeletal features for the i -th data instance. When heterogeneous skeletal features are used, each vector $\mathbf{x}_i \in \mathbb{R}^d$ consists of m modalities such that $d = \sum_{j=1}^m d_j$. Within each modality, the skeletal features are further divided into s partitions, and each partition contains features from a skeleton joint. Then, we formulate robot awareness of human behaviors as a problem of dividing $\{\mathbf{x}_i\}_{i=1}^n$ into c behavior categories through exploiting all available information from heterogeneous feature modalities and skeleton joints, using

a regression-like classification objective as follows:

$$\min_{\mathbf{W}} \|\mathbf{X}^\top \mathbf{W} + \mathbf{1}_n \mathbf{b}^\top - \mathbf{Y}\|_F^2, \quad (1)$$

where $\mathbf{1}_n \in \mathbb{R}^{n \times 1}$ is the constant vector of all 1's, $\mathbf{b} \in \mathbb{R}^{c \times 1}$ is the intercept vector, $\mathbf{Y} = [\mathbf{y}_1, \dots, \mathbf{y}_n]^\top \in \mathbb{R}^{n \times c}$ denotes the behavior category indicator matrix, and $\mathbf{y}_i \in \mathbb{R}^c$ denotes the category indicator vector for the feature vector \mathbf{x}_i with y_{ij} indicating how likely \mathbf{x}_i belongs to the j -th category. The label matrix \mathbf{Y} of the data instances is given in the training phase. Then, the value of \mathbf{b} in Eq. (1) can be calculated by $\mathbf{b} = \mathbf{Y}^\top \mathbf{1}_n / n$.

The solution of the optimization problem in Eq. (1) is the parameter matrix $\mathbf{W} = [\mathbf{w}_1, \mathbf{w}_2, \dots, \mathbf{w}_c] \in \mathbb{R}^{d \times c}$, which contains the weights $\mathbf{w}_i \in \mathbb{R}^d$ of each feature modality and skeletal joint with respect to the i -th behavior category. The parameter matrix \mathbf{W} is denoted as:

$$\mathbf{W} = \begin{bmatrix} \mathbf{w}_1^1 & \dots & \mathbf{w}_c^1 \\ \vdots & \ddots & \vdots \\ \mathbf{w}_1^m & \dots & \mathbf{w}_c^m \end{bmatrix}, \quad (2)$$

where $\mathbf{w}_p^q \in \mathbb{R}^{d_q}$ indicates the weights of the q -th modality including all skeleton joints with respect to the p -th behavior category, which is denoted as $\mathbf{w}_p^q = [\mathbf{w}_p^{q1}; \mathbf{w}_p^{q2}; \dots; \mathbf{w}_p^{qs}]$, and $\mathbf{w}_p^{qr} \in \mathbb{R}^{d_{qr}}$ represents the weights of the r -th skeleton joint within the q -th modality with respect to the p -th human behavior category, where d_{qr} is the dimension of features that are obtained from the r -th skeleton joint in the q -th modality, satisfying $\sum_{r=1}^s d_{qr} = d_q$, and s is the number of skeleton joints in each modality. An illustration of the weight matrix is presented in Fig. 2.

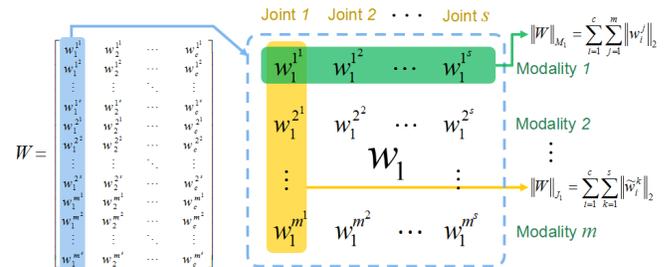


Fig. 2. Illustration of the structured sparsity-inducing norms introduced in our FABL method. Given the parameter matrix $\mathbf{W} = [\mathbf{w}_1, \mathbf{w}_2, \dots, \mathbf{w}_c] \in \mathbb{R}^{d \times c}$, we arrange each column vector \mathbf{w}_i of the i -th action category into a matrix, where rows represent modalities and columns denote skeletal joints. We model the interrelationships of the feature modalities using the M_1 -norm regularization term, and the interrelationships of the skeletal joints using the J_1 -norm regularization to model the representative joints.

B. Learning of Discriminative Body Parts

For specific behaviors, a small set of body parts (represented as joints in human skeletal data) are more discriminative than others. For example, in the behavior of hand waving as depicted in Fig. 1(b), the forehead and hand joints are more discriminative. Such discriminative human skeletal joints are typically not shared by all behavior categories (i.e. the joints to recognize waving and kicking are

substantially different). To learn discriminative body parts, we introduce a new joint-based group ℓ_1 -norm (named J_1 -norm) as a regularizer of the problem in Eq. (1). The J_1 -norm is mathematically defined as $\|\mathbf{W}\|_{J_1} = \sum_{i=1}^c \sum_{k=1}^s \|\tilde{\mathbf{w}}_i^k\|_2$, where $\tilde{\mathbf{w}}_i^k \in \mathbb{R}^{d_k}$ denotes the weights of the k -th human skeletal joint with respect to the i -th behavior category for all feature modalities, which is expressed as $\tilde{\mathbf{w}}_i^k = [\mathbf{w}_i^{1k}; \mathbf{w}_i^{2k}; \dots; \mathbf{w}_i^{mk}]$, and $\sum_{k=1}^s d_k = d$. Then, we can rewrite the objective function as:

$$\min_{\mathbf{W}} \|\mathbf{X}^\top \mathbf{W} + \mathbf{1}_n \mathbf{b}^\top - \mathbf{Y}\|_F^2 + \gamma \|\mathbf{W}\|_{J_1}. \quad (3)$$

where γ is a trade-off hyperparameter.

The J_1 -norm applies the ℓ_2 -norm within each skeletal joint and the ℓ_1 -norm between the joints, which enforces sparsity among different joints. For example, if the skeletal features obtained from a human skeleton joint are not discriminative for a specific behavior category, the objective in Eq. (3) will assign zeros (in the ideal case, usually very small values) to them for this behavior category; otherwise, their weights have large values. As shown in Fig. 2, the J_1 -norm regularization term captures the interrelationship among body parts, and estimates the importance of each body part to identify certain human behaviors.

C. Learning of Multimodal Skeletal Features

When heterogeneous multimodal features are available, it is well accepted that different types of skeletal features show varying performance on recognizing different behaviors [4]. That is, the features from a specific modality can be more or less discriminative for recognizing specific human behaviors. For example, comparing to pose features, motion features are generally less helpful to identify a still human behavior such as sitting. To integrate multiple feature modalities and model their interrelationships, we introduce another group ℓ_1 -norm (M_1 -norm) as a new regularizer in Eq. (3), which is defined as $\|\mathbf{W}\|_{M_1} = \sum_{i=1}^c \sum_{j=1}^m \|\mathbf{w}_i^j\|_2$. Then, incorporating both multi-feature and multi-joint group sparsity-inducing norms, the final objective function becomes:

$$\min_{\mathbf{W}} \|\mathbf{X}^\top \mathbf{W} + \mathbf{1}_n \mathbf{b}^\top - \mathbf{Y}\|_F^2 + \gamma_1 \|\mathbf{W}\|_{M_1} + \gamma_2 \|\mathbf{W}\|_{J_1}. \quad (4)$$

where γ_1 and γ_2 are trade-off hyperparameters.

The M_1 -norm uses the ℓ_2 -norm within each feature modality and the ℓ_1 -norm between these modalities, which enforces the sparsity of these modalities. For example, if a modality is not discriminative enough to recognize a certain behavior category, the objective in Eq. (4) will assign zeros (in the ideal case, usually very small values) to the features within this modality with respect to the behavior category; otherwise, their weights are large. As demonstrated in Fig 2., the proposed M_1 -norm regularization term captures the interrelationship between feature modalities and estimates their importance to recognize certain behaviors.

D. Human Behavior Understanding

After solving the optimization problem in Eq. (4) during the training phase (solution is detailed in Section IV), we can

obtain the optimal weight matrix $\mathbf{W}^* = [\mathbf{w}_1^*, \mathbf{w}_2^*, \dots, \mathbf{w}_c^*] \in \mathbb{R}^{d \times c}$. Then, in the testing phase, given a new multisensory instance $\mathbf{x} \in \mathbb{R}^d$, its behavior category $y(\mathbf{x})$ is decided by:

$$y(\mathbf{x}) = \underset{i}{\operatorname{argmax}} \mathbf{x}^\top \mathbf{w}_i^* + b_i, \quad i = 1, 2, \dots, c. \quad (5)$$

An advantage of our formulation utilizing the regression-like objective function is that classification is integrated with feature learning; thus, we do not require additional classifiers (e.g., SVMs). This significantly improves processing efficiency, resulting in high-speed recognition of human behaviors that can benefit real-time human-centered robotics applications.

IV. OPTIMIZATION ALGORITHM

Since the objective in Eq. (4) comprises two non-smooth regularization terms: the M_1 -norm and J_1 -norm, it is difficult to solve in general. To this end, we implement a new iterative algorithm to solve the optimization problem in Eq. (4) with non-smooth regularization terms. The proposed optimization solver has a theoretical guarantee to find the optimal solution.

To learn the value of the weight matrix \mathbf{W} , we compute the derivative of the objective with respect to \mathbf{w}_i ($1 \leq i \leq c$) and set it to zero vector. Then, we obtain

$$\mathbf{X}\mathbf{X}^\top \mathbf{w}_i - \mathbf{X}(\mathbf{y}_i - \mathbf{b}_i) + \gamma_1 \mathbf{D}^i \mathbf{w}_i + \gamma_2 \tilde{\mathbf{D}}^i \mathbf{w}_i = \mathbf{0}, \quad (6)$$

where \mathbf{D}^i ($1 \leq i \leq c$) is a block diagonal matrix with the j -th diagonal block as $\frac{1}{2\|\mathbf{w}_i^j\|_2} \mathbf{I}_j$, \mathbf{w}_i^j is the j -th segment of \mathbf{w}_i consisting of the weights of the j -th feature, $\tilde{\mathbf{D}}^i$ is a diagonal matrix with the k -th diagonal block as $\frac{1}{2\|\tilde{\mathbf{w}}_i^k\|_2} \mathbf{I}_k$, $\tilde{\mathbf{w}}_i^k$ is the k -th segment of \mathbf{w}_i including the weights of skeletal features calculated from the k -th skeleton joint, and \mathbf{I}_j is the identity matrix of size d_j . Thus we have

$$\mathbf{w}_i = (\mathbf{X}\mathbf{X}^\top + \gamma_1 \mathbf{D}^i + \gamma_2 \tilde{\mathbf{D}}^i)^{-1} \mathbf{X}(\mathbf{y}_i - \mathbf{b}_i). \quad (7)$$

Both \mathbf{D}^i and $\tilde{\mathbf{D}}^i$ are dependent on \mathbf{W} and thus also unknown variables. An iterative algorithm is implemented to solve this problem, which is described in Algorithm 1.

Before analyzing convergence of Algorithm 1, we describe a lemma from [25] as follows.

Lemma 1: Given vectors \mathbf{a} and \mathbf{b} , the following equation holds

$$\|\mathbf{a}\|_2 - \frac{\|\mathbf{a}\|_2^2}{2\|\mathbf{b}\|_2} \leq \|\mathbf{b}\|_2 - \frac{\|\mathbf{b}\|_2^2}{2\|\mathbf{b}\|_2} \quad (8)$$

Theorem 1: Algorithm 1 converges to the optimal solution to the optimization problem in Eq. (4).

Proof: According to Step 3 of Algorithm 1, we know

$$\begin{aligned} \mathbf{W}(t+1) = \underset{\mathbf{W}}{\operatorname{argmin}} & \|\mathbf{X}^\top \mathbf{W} + \mathbf{1}_n \mathbf{b}^\top - \mathbf{Y}\|_F^2 \\ & + \gamma_1 \sum_{i=1}^c \mathbf{w}_i^\top \mathbf{D}^i(t+1) \mathbf{w}_i + \gamma_2 \sum_{i=1}^c \tilde{\mathbf{w}}_i^\top \tilde{\mathbf{D}}^i(t+1) \tilde{\mathbf{w}}_i. \end{aligned} \quad (9)$$

Then, we can derive that

$$\begin{aligned}
& \mathcal{J}(t+1) + \gamma_1 \sum_{i=1}^c \mathbf{w}_i^\top(t+1) \mathbf{D}^i(t+1) \mathbf{w}_i(t+1) \\
& + \gamma_2 \sum_{i=1}^c \tilde{\mathbf{w}}_i^\top(t+1) \tilde{\mathbf{D}}^i(t+1) \tilde{\mathbf{w}}_i(t+1) \\
\leq & \mathcal{J}(t) + \gamma_1 \sum_{i=1}^c \mathbf{w}_i^\top(t) \mathbf{D}^i(t+1) \mathbf{w}_i(t) \\
& + \gamma_2 \sum_{i=1}^c \tilde{\mathbf{w}}_i^\top(t) \tilde{\mathbf{D}}^i(t+1) \tilde{\mathbf{w}}_i(t), \tag{10}
\end{aligned}$$

where $\mathcal{J}(t) = \|\mathbf{X}^\top \mathbf{W}(t) + \mathbf{1}_n \mathbf{b}^\top - \mathbf{Y}\|_F^2$.

After substituting the definition of \mathbf{D}^i and $\tilde{\mathbf{D}}^i$, we obtain

$$\begin{aligned}
& \mathcal{J}(t+1) + \gamma_1 \sum_{i=1}^c \sum_{j=1}^m \frac{\|\mathbf{w}_i^j(t+1)\|_2^2}{2\|\mathbf{w}_i^j(t)\|_2} \\
& + \gamma_2 \sum_{i=1}^c \sum_{k=1}^s \frac{\|\tilde{\mathbf{w}}_i^k(t+1)\|_2^2}{2\|\tilde{\mathbf{w}}_i^k(t)\|_2} \\
\leq & \mathcal{J}(t) + \gamma_1 \sum_{i=1}^c \sum_{j=1}^m \frac{\|\mathbf{w}_i^j(t)\|_2^2}{2\|\mathbf{w}_i^j(t)\|_2} \\
& + \gamma_2 \sum_{i=1}^c \sum_{k=1}^s \frac{\|\tilde{\mathbf{w}}_i^k(t)\|_2^2}{2\|\tilde{\mathbf{w}}_i^k(t)\|_2}. \tag{11}
\end{aligned}$$

From Lemma 1, we can derive

$$\begin{aligned}
& \sum_{j=1}^m \|\mathbf{w}_i^j(t+1)\|_2 - \sum_{j=1}^m \frac{\|\mathbf{w}_i^j(t+1)\|_2^2}{2\|\mathbf{w}_i^j(t)\|_2} \leq \\
& \sum_{j=1}^m \|\mathbf{w}_i^j(t)\|_2 - \sum_{j=1}^m \frac{\|\mathbf{w}_i^j(t)\|_2^2}{2\|\mathbf{w}_i^j(t)\|_2}, \tag{12}
\end{aligned}$$

and

$$\begin{aligned}
& \sum_{k=1}^s \|\tilde{\mathbf{w}}_i^k(t+1)\|_2 - \sum_{k=1}^s \frac{\|\tilde{\mathbf{w}}_i^k(t+1)\|_2^2}{2\|\tilde{\mathbf{w}}_i^k(t)\|_2} \leq \\
& \sum_{k=1}^s \|\tilde{\mathbf{w}}_i^k(t)\|_2 - \sum_{k=1}^s \frac{\|\tilde{\mathbf{w}}_i^k(t)\|_2^2}{2\|\tilde{\mathbf{w}}_i^k(t)\|_2}. \tag{13}
\end{aligned}$$

Adding Eqs. (11)-(13) on both sides, we obtain

$$\begin{aligned}
& \mathcal{J}(t+1) + \gamma_1 \sum_{i=1}^c \sum_{j=1}^m \|\mathbf{w}_i^j(t+1)\|_2 \\
& + \gamma_2 \sum_{i=1}^c \sum_{k=1}^s \|\tilde{\mathbf{w}}_i^k(t+1)\|_2 \\
\leq & \mathcal{J}(t) + \gamma_1 \sum_{i=1}^c \sum_{j=1}^m \|\mathbf{w}_i^j(t)\|_2 + \gamma_2 \sum_{i=1}^c \sum_{k=1}^s \|\tilde{\mathbf{w}}_i^k(t)\|_2. \tag{14}
\end{aligned}$$

Therefore, Algorithm 1 decreases the objective value in each iteration. Since the optimization problem defined in Eq. (4) is convex, and the objective is lower-bounded by zero due to the definition of matrix and vector norms, thus the algorithm converges to the optimum. ■

Algorithm 1: An iterative algorithm to solve the problem in Eq. (4)

Input : $\mathbf{X} = [\mathbf{x}_1, \dots, \mathbf{x}_n] \in \mathbb{R}^{d \times n}$ and $\mathbf{Y} = [\mathbf{y}_1, \dots, \mathbf{y}_n]^\top \in \mathbb{R}^{n \times c}$

- 1 Let $t = 1$. Initialize $\mathbf{W}(t)$ by solving $\min_{\mathbf{W}} \|\mathbf{X}^\top \mathbf{W} + \mathbf{1}_n \mathbf{b}^\top - \mathbf{Y}\|_F^2$.
- 2 **while not converge do**
- 3 Calculate the block diagonal matrix $\mathbf{D}^i(t+1)$ ($1 \leq i \leq c$), where the j -th diagonal block of $\mathbf{D}^i(t+1)$ is $\frac{1}{2\|\mathbf{w}_i^j(t)\|_2} \mathbf{I}_j$.
 Calculate the block diagonal matrix $\tilde{\mathbf{D}}^i(t+1)$ ($1 \leq i \leq c$), where the k -th diagonal block of $\tilde{\mathbf{D}}^i(t+1)$ is $\frac{1}{2\|\tilde{\mathbf{w}}_i^k(t)\|_2} \mathbf{I}_k$.
- 4 For each \mathbf{w}_i ($1 \leq i \leq c$), $\mathbf{w}_i(t+1) = (\mathbf{X}\mathbf{X}^\top + \gamma_1 \mathbf{D}^i(t+1) + \gamma_2 \tilde{\mathbf{D}}^i(t+1))^{-1} \mathbf{X}(\mathbf{y}_i - \mathbf{b}_i)$.
- 5 $t = t + 1$.

Output: $\mathbf{W} = \mathbf{W}(t) \in \mathbb{R}^{d \times c}$

V. EXPERIMENTS

To quantitatively assess the performance of the proposed FABL method, we conduct experiments using public benchmark datasets. Furthermore, to evaluate the benefits of our FABL method in real-world robotics applications, we deploy FABL on a Baxter robot to perform online, real-time behavior recognition for human-robot interaction.

A. Implementation

Our FABL approach is implemented using a combination of Matlab and C++ on a Linux machine with an i7 3.4GHz CPU and 16GB memory. The Matlab code is used to validate our approach on two public datasets: MSR Action3D Dataset [26] and Cornell Activity Dataset [6], while the C++ program is employed for validation on a Baxter robot in a real-world ‘‘serving drinks’’ task.

We intentionally designed and applied four simple skeletal features to emphasize the performance gain resulted from our FABL method instead of sophisticated features. These simple skeletal features include: (1) spatial joint displacement that is the 3D coordinate difference of each body part with respect to the torso: $(\Delta x, \Delta y, \Delta z) = (x, y, z) - (x^c, y^c, z^c)$, where (x, y, z) represents the coordinates of each skeletal joint, and (x^c, y^c, z^c) denotes the coordinates of the center torso joint in skeletal data, (2) temporal joint displacement, which is defined as the temporal location difference of the same body joint in the current frame with respect to the previous frame: $(\dot{x}, \dot{y}, \dot{z}) = (x_t, y_t, z_t) - (x_{t-1}, y_{t-1}, z_{t-1})$, where (x_t, y_t, z_t) is the joint location at time t , (3) long-term temporal joint displacement, defined as the temporal 3D location difference between the current frame and the initial frame: $(\ddot{x}, \ddot{y}, \ddot{z}) = (x_t, y_t, z_t) - (x_0, y_0, z_0)$, where (x_0, y_0, z_0) is the coordinates of a joint in the initial frame, and (4) spatial joint distance, which is defined as the geometrical distance of a joint to the torso center joint: $d = \|(x, y, z) - (x^c, y^c, z^c)\|_2$. Then, we

compute a histogram of each feature type to build a vector that is used as a feature modality in our experiment.

B. Results on MSR Action3D Dataset

We evaluate the performance of the proposed approach to recognize human behaviors when interacting with structured-light cameras, using the MSR Action3D benchmark dataset [26]. This dataset contains 20 categories of human actions performed by 7 subjects for three times. The skeleton sequence of “high arm waving” is shown in Fig. 3.

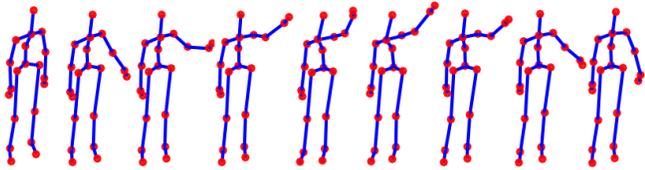


Fig. 3. The MSR Action3D dataset is utilized in the experiment to evaluate the proposed FABL approach, which contains 20 activities recorded using Kinect, which are (M1) high arm wave, (M2) horizontal arm wave, (M3) hammer, (M4) hand catch, (M5) forward punch, (M6) high throw, (M7) draw x, (M8) draw tick, (M9) draw circle, (M10) hand clap, (M11) two hand wave, (M12) side boxing, (M13) bend, (M14) forward kick, (M15) side kick, (M16) jogging, (M17) tennis swing, (M18) tennis serve, (M19) golf swing, and (M20) pick up & throw. This figure shows a sample skeleton sequence of the action (M1) high arm waving in the dataset

We evaluate the recognition performance using a challenging subject-wise setting. That is, the training dataset does not contain any data instances from the subjects who participate in testing. When combined both structured sparsity-inducing norms to perform simultaneous feature and skeletal joint learning, our FABL method obtains an accuracy of 91.67%, The confusion matrix obtained by our method is shown in Fig. 4(a), which demonstrates our FABL approach is able to well recognize most of the behaviors. The actions that are not well identified is (M4) hand-catch, and (M7) draw-x that is always misclassified as the action of (M8) draw tick or (M9) draw circle, which have similar, small motions.

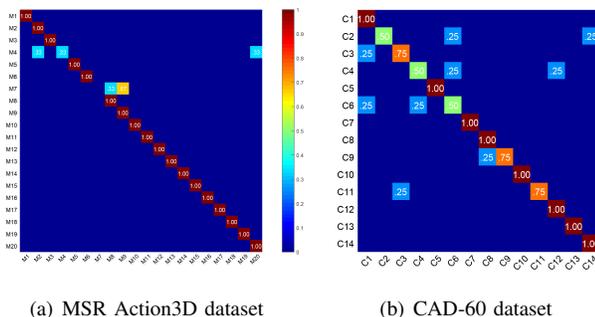


Fig. 4. Confusion matrices obtained by our FABL method over the MSR Action3D and CAD-60 dataset datasets. The behavior category labels *M1-M16* and *C1-C14* are described in Fig. 3 and Fig. 5, respectively.

We compare with two baseline methods including feature-learning-only ($\gamma_2 = 0$) and body-part-learning-only ($\gamma_1 = 0$). As presented in Table I, the feature-learning-only method obtains an average recognition accuracy of 85.00%, while

the body-part-learning-only obtains an average accuracy of 86.67%. This indicates that FABL outperforms baseline approaches using a single norm for regularization. In addition, we compare our FABL method with previous activity recognition techniques based on skeleton features. FABL achieves promising recognition accuracy (with the high-speed performance) on the MSR Action3D dataset.

TABLE I
COMPARISON OF AVERAGE ACCURACY WITH PREVIOUS SKELETON-BASED METHODS ON THE MSR ACTION3D DATASET

Reference	Method	Accuracy
Ofli et al. [11]	Sequence of Most Informative Joints	41.18%
Wang et al. [10]	Dynamic Temporal Warping	54.0%
Ellis et al. [27]	Joints Distance + Key Poses	65.7%
Li et al. [26]	Action Graph	74.7%
Xia et al. [28]	HOJ3D	78%
Yang and Tian [9]	EigenJoints	83.3%
Wang et al. [2]	Actionlet Ensemble	88.2%
Ben Amor et al. [29]	Skeleton Trajectories	89%
Our Methods	Feature Learning Only	85.00%
	Body-Part Learning Only	86.67%
	FABL	91.67%

C. Results on Cornell Activity Dataset

The Cornell Activity Dataset 60 (CAD-60) [6] is a widely applied benchmark for human activity recognition in robotics applications. This dataset includes color-depth and skeleton information of twelve daily activities as well as two motions “still” and “random” recorded by a Kinect sensor in various environments, including office, kitchen, bedroom, bathroom, and living room. Each activity is performed by four subjects with two males and two females (one subject is left-handed). The skeleton data in each frame contains 15 joints, as shown in Figure 5. We evaluate FABL’s performance in a subject-wise cross-validation setup [30], where actions performed by new subjects are used for testing.

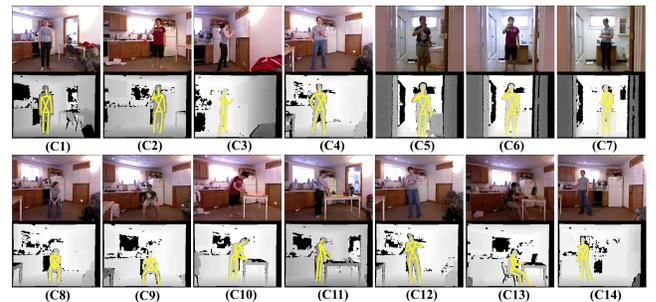


Fig. 5. The CAD-60 dataset contains 14 behaviors, including (C1) standing still, (C2) talking on the phone, (C3) writing on whiteboard, (C4) drinking water, (C5) rinsing mouth with water, (C6) brushing teeth, (C7) wearing contact lenses, (C8) talking on couch, (C9) relaxing on couch, (C10) cooking (chopping), (C11) cooking (stirring), (C12) opening pill container, (C13) working on computer, (C14) random. RGB images are depicted in the top row, and the depth images with the human skeleton in yellow are shown in the bottom row.

As demonstrated in Table II, the FABL method using both regularization terms obtain an average accuracy of 83.93%, and its detailed confusion matrix is graphically presented in

Fig. 4(b), which generally indicates that most of the activities can be well classified by our approach.

TABLE II

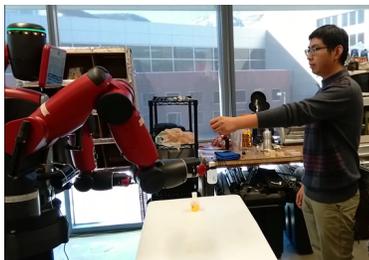
COMPARISON OF AVERAGE RECOGNITION ACCURACY WITH PREVIOUS SKELETON-BASED METHODS ON THE CAD-60 DATASET

Reference	Method	Accuracy
Ni <i>et al.</i> [30]	Order-Preserving Sparse Coding	65.32%
Piyathilaka and Kodagoda [31]	Hidden Markov Model	78.38%
Wang <i>et al.</i> [2]	Skeleton-based Actionlet Ensemble	74.70%
Zhang and Tian [32]	Bag of Features	80.77%
Our Methods	Feature Learning Only	78.57%
	Body-Part Learning Only	79.46%
	FABL	83.93%

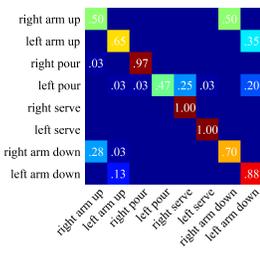
We implemented two baseline techniques under the same formulation. First, we set $\gamma_2 = 0$ to evaluate the performance of the feature learning scheme, and obtain an accuracy of 78.57%. Then, γ_1 is set to zero to evaluate the performance of the body-part learning scheme, and we obtain an average accuracy of 79.46%. It is observed that both baseline methods perform worse than the full FABL approach using both regularization terms. Moreover, we implemented a third baseline method with no regularization terms, which obtains an accuracy of 76.79% and performs worse than the methods with the regularization terms. In addition, we compare our FABL method with previous state-of-the-art skeleton-based techniques for activity recognition, as reported in Table II, which shows our FABL method outperforms these skeleton-based techniques over the CAD-60 dataset.

D. Behavior Recognition for Human-Robot Interaction

Besides using public benchmark datasets to evaluate and compare our FABL method’s accuracy, we also implemented and deployed the method on a physical robot to validate its performance in real-world robotics applications. The robot employed in this experiment is a Baxter robot, as shown in Fig. 6(a), which uses a structured-light sensor for onboard 3D perception and the same workstation (Intel i7 3.4GHz CPU and 16GB memory) for onboard control and data processing.



(a) Baxter performing “serving drinks”



(b) Confusion matrix

Fig. 6. We evaluate our FABL approach using a Baxter robot to recognize behaviors for real-time human-robot interaction. The tasks focus on the robot-assisted living application such as “serving drinks” as shown in Fig. 6(a). The confusion matrix is illustrated in Fig. 6(b).

In this experiment, the task focuses on the robot-assisted living application, where the Baxter robot needs to recognize

the activities of a subject and perform a collection of predefined robot actions, such as “serving drinks,” as demonstrated in Fig. 6(a), in response to the subject’s activity. We define six robot actions, including fetching a drinking bottle with one hand, fetching an empty cup with the other hand, pouring the drinks into the cup, putting back the bottle, serving the drinking cup to the subject, and finally putting back the cup. Each robot action is triggered by a specific command gesture performed by a subject in front of the robot, which must be recognized by the Baxter robot. The skeleton data is captured onboard and in real time using ROS and the OpenNI package.

Eight human behavior categories are defined and used to interact with the robot, including lifting up left/right arms, pouring with left/right hands, serving with left/right hands, and putting down left/right arms. We specifically distinguish between left side and right side, because this is critical to take into account human preference in practical, real-world scenarios. Two human subjects having different body scales and motion patterns are involved in this experiment. Each subject performs each of the eight behaviors 20 times. Actions by one subject were used for training, while other subject’s actions were used for testing. Ground truth is manually recorded and used to compare with recognition results obtained by the robot for quantitative evaluation. After extracting multimodal features from training instances, our method computes the optimal weight matrix by Algorithm 1 using the training data. Then, the learned FABL approach is deployed on the robot for online, onboard behavior recognition to enable real-time human-robot interaction.

Similar to the experiments using public datasets, we also quantitatively assess FABL’s performance and compare with baseline and existing skeleton-based techniques. The average accuracy obtained by the complete FABL method is 77.19% with both regularization terms. The confusion matrix obtained by our FABL approach is demonstrated in Fig. 6(b). For comparison, the baseline technique based only on feature learning ($\gamma_2 = 0$) obtains an accuracy of 76.56%, while the baseline based only on body part learning ($\gamma_1 = 0$) obtains an average recognition accuracy of 76.25%. In addition, we compare our FABL method with several previous skeleton-based recognition techniques and present the results in Table III. We can observe that the FABL method is able to obtain better performance over baseline and used previous methods. Since only one subject’s actions were used to train the FABL model, the recognition accuracy was not as significant as that using public benchmark datasets. More training data will improve the testing performance.

VI. DISCUSSION

High-Speed Processing. Due to the capability of our FABL approach to integrate both feature learning and classification in the same formulation, and the efficiency of our regression-like objective function, our FABL approach is able to achieve high-speed processing. To validate this strong advantage, we perform additional experiments over the MSR Action3D and CAD-60 datasets using Matlab implementations without any optimization, and utilizing the real Baxter robot using a C++

TABLE III

COMPARISON OF AVERAGE RECOGNITION ACCURACY WITH PREVIOUS METHODS FOR REAL-TIME HUMAN-ROBOT INTERACTION

Reference	Method	Accuracy
[27]	Relative Angles and Distances	15.00%
[2]	Histogram of Joint Position Differences	48.13%
[8]	Histogram of Oriented Displacements	51.25%
Our Methods	Feature Learning Only	76.56%
	Body-Part Learning Only	76.25%
	FABL	77.19%

implementation. The runtime results on all used datasets are presented in Table IV, which shows our FABL approach can achieve a significantly high processing speed at the order of 10^4 Hz. This indicates the promise of our FABL approach to identify human behaviors in real-time robotics applications.

TABLE IV

RUNTIME ANALYSIS OVER DIFFERENT DATASETS

Runtime	MSR Action3D	CAD-60	Baxter
Processing speed (Hz)	2.2×10^4	1.4×10^4	3.3×10^4
Time per observation (s)	4.5×10^{-5}	7.3×10^{-5}	3.0×10^{-5}

Generalizability. FABL is a general approach that can work with different body kinematic models obtained by a variety of sensing devices and skeleton generation packages, including the OpenNI package in ROS, Microsoft SDKs, and MoCap systems. Given any kinematic body model from the devices, we can downsample the body model into 15 body parts, and apply FABL to automatically identify the most representative parts. In this case, FABL can achieve cross-training [5], i.e., methods trained on a kinematic body model from one device can be directly applied to other models by a different device, which can significantly save design labor.

Hyperparameter Selection. The regularization hyperparameters γ_1 and γ_2 are utilized to control the effect of feature learning and the strength of body-part learning, respectively. Their optimal values can be decided using cross-validation during the training process. In general, we observe that the values $\gamma_1 = 0.1$ and $\gamma_2 = 0.1$ usually result in satisfactory recognition accuracy, which shows that both regularization terms are necessary. When the values of hyperparameters become too large, the performance decreases, because the loss function that models the recognition error is more ignored. When γ_1 and γ_2 take too small values, the approach cannot well capture the interrelationships of feature modalities and body parts, thus decreasing the recognition accuracy.

VII. CONCLUSION

In this paper, we introduce a novel FABL approach that is able to simultaneously learn discriminative feature modalities and body parts to perform high-speed human behavior recognition. The proposed FABL method automatically identifies discriminative feature modalities and important body parts using two structured sparsity-inducing norms to model their interrelationships. Our FABL approach formulates behavior recognition as a regression-like optimization problem, which

is solved by an efficient iteration algorithm that possesses a theoretical guarantee to find the optimal solution. To evaluate the performance of the proposed FABL method, we perform empirical studies using two public benchmark datasets and a physical Baxter robot. The experimental results have indicated that FABL is able to outperform existing skeleton-based methods. More importantly, our FABL approach achieves a high processing speed of more than 10^4 Hz, which can enable realistic, self-contained, intelligent robots to recognize human behaviors and interact with humans in real time.

REFERENCES

- [1] H. Zhang, W. Zhou, C. Reardon, and L. E. Parker, "Simplex-based 3D spatio-temporal feature description for action recognition," in *CVPR*, 2014.
- [2] J. Wang, Z. Liu, Y. Wu, and J. Yuan, "Learning actionlet ensemble for 3D human action recognition," *TPAMI*, vol. 36, no. 5, pp. 914–927, 2014.
- [3] J. Shotton, A. Fitzgibbon, M. Cook, T. Sharp, M. Finocchio, R. Moore, A. Kipman, and A. Blake, "Real-time human pose recognition in parts from single depth images," in *CVPR*, 2011.
- [4] F. Han, B. Reily, W. Hoff, and H. Zhang, "Space-time representation of people based on 3D skeletal data: A review," *CVIU*, 2017, to appear.
- [5] H. Zhang and L. E. Parker, "Bio-inspired predictive orientation decomposition of skeleton trajectories for real-time human activity prediction," in *ICRA*, 2015.
- [6] J. Sung, C. Ponce, B. Selman, and A. Saxena, "Unstructured human activity detection from RGBD images," in *ICRA*, 2012.
- [7] X. Yang and Y. Tian, "EigenJoints-based action recognition using Naïve-Bayes-Nearest-Neighbor," in *CVPRW*, 2012.
- [8] M. A. Gowayyed, M. Torki, M. E. Hussein, and M. El-Saban, "Histogram of oriented displacements (HOD): describing trajectories of human joints for action recognition," in *IJCAI*, 2013.
- [9] X. Yang and Y. Tian, "Effective 3D action recognition using Eigen-Joints," *JVCIR*, vol. 25, no. 1, pp. 2–11, 2014.
- [10] J. Wang, Z. Liu, Y. Wu, and J. Yuan, "Mining actionlet ensemble for action recognition with depth cameras," in *CVPR*, 2012.
- [11] F. Ofli, R. Chaudhry, G. Kurillo, R. Vidal, and R. Bajcsy, "Sequence of the most informative joints (SMIJ): A new representation for human skeletal action recognition," *JVCIR*, vol. 25, no. 1, pp. 24–38, 2014.
- [12] X. Chen and M. Koskela, "Online RGB-D gesture recognition with extreme learning machines," in *ICMI*, 2013.
- [13] P. Wei, Y. Zhao, N. Zheng, and S.-C. Zhu, "Modeling 4D human-object interactions for event and object recognition," in *ICCV*, 2013.
- [14] P. Wei, N. Zheng, Y. Zhao, and S.-C. Zhu, "Concurrent action detection with structural prediction," in *AAAI*, 2013.
- [15] G. Yu, Z. Liu, and J. Yuan, "Discriminative orderlet mining for real-time recognition of human-object interaction," in *ACCV*, 2014.
- [16] S. Z. Masood, C. Ellis, A. Nagaraja, M. F. Tappen, J. J. L. Jr., and R. Sukthankar, "Measuring and reducing observational latency when recognizing actions," in *ICCV*, 2011.
- [17] C. Wang, Y. Wang, and A. L. Yuille, "An approach to pose-based action recognition," in *CVPR*, 2013.
- [18] B. X. Nie, C. Xiong, and S.-C. Zhu, "Joint action recognition and pose estimation from video," in *CVPR*, 2015.
- [19] Y. Du, W. Wang, and L. Wang, "Hierarchical recurrent neural network for skeleton based action recognition," in *CVPR*, 2015.
- [20] A. A. Chaaoui, J. R. Padilla-López, P. Climent-Pérez, and F. Flórez-Revuelta, "Evolutionary joint selection to improve human action recognition with RGB-D devices," *ESA*, vol. 41, no. 3, pp. 786–794, 2014.
- [21] M. Reyes, G. Domínguez, and S. Escalera, "Feature weighting in dynamic timewarping for gesture recognition in depth data," in *ICCVW*, 2011.
- [22] O. Patsadu, C. Nukoolkit, and B. Watanapa, "Human gesture recognition using Kinect camera," in *IJCSSE*, 2012.
- [23] D.-A. Huang and K. M. Kitani, "Action-reaction: Forecasting the dynamics of human interaction," in *ECCV*, 2014.

- [24] R. Chaudhry, F. Ofli, G. Kurillo, R. Bajcsy, and R. Vidal, "Bio-inspired dynamic 3D discriminative skeletal features for human action recognition," in *CVPR*, 2013.
- [25] F. Nie, H. Huang, X. Cai, and C. H. Ding, "Efficient and robust feature selection via joint $\ell_{2,1}$ -norms minimization," in *NIPS*, 2010.
- [26] W. Li, Z. Zhang, and Z. Liu, "Action recognition based on a bag of 3D points," in *CVPRW*, 2010.
- [27] C. Ellis, S. Z. Masood, M. F. Tappen, J. J. Laviola Jr, and R. Suktanar, "Exploring the trade-off between accuracy and observational latency in action recognition," *IJCV*, vol. 101, no. 3, pp. 420–436, 2013.
- [28] L. Xia, C.-C. Chen, and J. Aggarwal, "View invariant human action recognition using histograms of 3D joints," in *CVPRW*, 2012.
- [29] B. Ben Amor, J. Su, and A. Srivastava, "Action recognition using rate-invariant analysis of skeletal shape trajectories," *TPAMI*, vol. 38, no. 1, pp. 1–13, 2016.
- [30] B. Ni, P. Moulin, and S. Yan, "Order-preserving sparse coding for sequence classification," in *ECCV*, 2012.
- [31] L. Piyathilaka and S. Kodagoda, "Gaussian mixture based HMM for human daily activity recognition using 3D skeleton features," in *CIEA*, 2013.
- [32] C. Zhang and Y. Tian, "RGB-D camera-based daily living activity recognition," *CVIP*, vol. 2, no. 4, p. 12, 2012.