Robust Real-Time Group Activity Recognition of Robot Teams

Lyujian Lu, Hua Wang, Brian Reily and Hao Zhang

***Abstract*—Recognition of group activities is critical for the success of applications that depend on effective human-robot teaming. Awareness of these group activities (also referred to team behaviors in some literature), including the individual activities of human teammates and the overall team intent, allows robotic teammates to work alongside humans without explicit commands and to offer proactive assistance towards the overall mission. In this paper, we present a novel approach to robot recognition of team activities, simultaneously learning a projection from multi-sensory input data to a latent representation of individual activities and a projection from this representation to the overall activities. We introduce a smoothed iterative reweighted algorithm to solve this formulated optimization problem, guaranteed to converge to an optimal solution. We evaluate our approach extensively on benchmark group and team activity datasets, showing that our approach achieves state of the art performance while operating in real-time on mobile robots.**

***Index Terms*—Group Activity Recognition, Robot Teaming, Real Time Recognition**

I. Introduction

EFFECTIVE human-robot teaming is critical problem when humans and robots must work alongside each other to achieve common goals. In a large number of real world applications where time and safety are paramount, such as search and rescue and disaster response, humans and robots must act as a team, working towards a common goal while performing their own individual tasks. Real-world environments where these missions are performed are often hazardous, making it dangerous and even impossible for commands and intents to be expressed, requiring that these common goals must be understood without explicit communications. In addition, rescuers and first responders may not be trained to interact with a robotic teammate, which requires robots to be able to understand their human teammates just as other humans would.

As autonomous robots are being increasingly integrated into human teams to perform tasks in dangerous and hazardous environments [1], robots need to intelligently and automatically recognize the activities of their human teammates (including the overall team activities and individual activities) in order to provide proactive support on an operation or directly assist with the overall mission goal, without cognitively burdening their human peers [2]. To achieve this, it is required that the robots can effectively recognize team activities and identify not just an individual's activity but the overall activity of the team.

Due to its importance, activity recognition has been extensively studied, which, though, mostly focused on modeling and recognizing the activities of individual humans [3], but not teams as a whole. Other approaches have been developed to address the problem of group activity recognition (also referred to team behaviors in some literature), where a collection of individuals perform the same action (*e.g*, dancing).

These have included hierarchical models [4], custom engineered features [5], and deep learning based approaches utilizing long short term memories [6] and recurrent neural networks [7]. While these approaches have had success at group activity recognition, they fail to address the problem of team activity recognition, where the overall shared goal may be distinct from individual actions.

In this paper, we present a novel approach to address the problem of team activity recognition, operating in real-time from multisensory data. We formulate team activity recognition as a regularized optimization problem, simultaneously learning a projection from multisensory input data to a latent representation of individual activities and learning a projection from this latent representation to the overall team activity, where the individual activities are modeled as latent variables, as illustrated in Fig. 1. In addition, our proposed approach explore the relationship of the members within a team globally and locally. While the low-rank regularization is imposed to discover the common task among different teammates, the Laplacian embedding is leveraged to preserve pairwise relation between teammates. Moreover, to better model the noisy environments of real-world applications using robot teams, ℓ_p-norm $(0 < p \leq 2)$ is utilized to substitute squared ℓ_2-norm for better robustness [8], [9], [10], [11]. Additional regularization terms can also be integrated into our approach to fuse multisensory inputs in the same unified formulation. Despite its nice proprieties of our proposed model, it is a non-smooth objective and difficult to solve in general. An efficient smoothed iteratively reweighed algorithm is proposed to solve the optimization problem. Extensive experiment results on two benchmark datasets have shown that our approach achieves superior accuracy and real-time performance.

Manuscript received: October 15, 2020; Revised January 8, 2021; Accepted January 31, 2021.

This paper was recommended for publication by Editor Jingang Yi upon evaluation of the Associate Editor and Reviewers' comments. The work of L. Lu and H. Wang was supported in part by the National Science Foundation (NSF) under Grant IIS 1652943, IIS 1849359, CNS 1932482, and CCF 2029543. The work of B. Reily and H. Zhang was supported in part by the NSF under Grant CNS 1823245 and IIS 1849359. (Corresponding author: Hua Wang.)

L. Lu, H. Wang, B. Reily and H. Zhang are with the Department of Computer Science, Colorado School of Mines, Golden, CO 80401, USA (email: lyujianlu@mymail.mines.edu; huawangcs@gmail.com; breily@mines.edu; hzhang@mines.edu).

Digital Object Identifier (DOI): see top of this page.

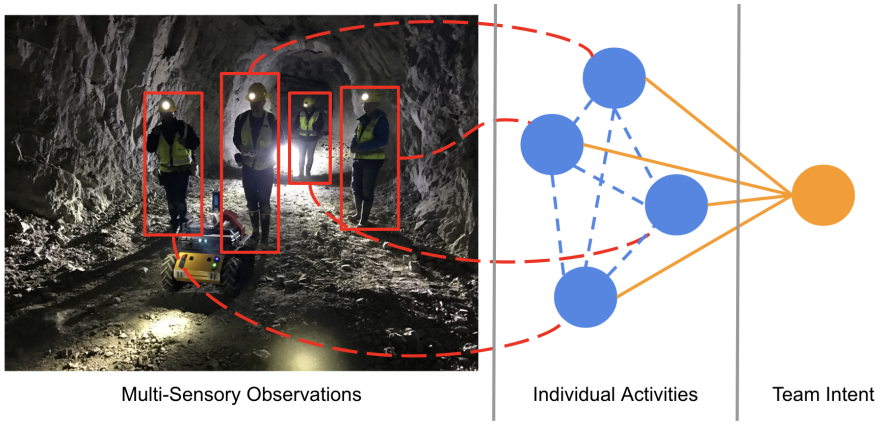


Fig. 1. In real-world scenarios, such as mine search and rescue, teammates perform individual tasks as they work towards a common goal. Our approach processes multisensory observations, learning latent representations of individual activities in order to recognize the overall team intent.

II. Related Work

Human-robot teaming is most successful when robots have an awareness of the human teammates' workload [12] and cognitive load [2]. Managing the cognitive load and building trust between robot and human teammates is critical for successful teaming [13], as humans prefer to working along autonomous and proactive robotic teammates [14]. As the key existing limitation to enabling seamless interaction to build trust and limit excess cognitive load is the ability of robots to understand the overall behavior of the team [15], recognition of individual and group activities is critical for team behavior understanding.

A. Individual Activity Recognition

Individual activity recognition approaches have often been based on skeletal data or visual observations of the individual [3]. Low-level features have been used, from discrete local features [16] to combinations of features [17]. Higher level features have also been applied, from engineered ones such as representing the movements of joints through anatomical planes [18] to learned features based on identifying discriminative joints and sensing modalities through regularized optimization. Deep learning has also been applied recently, particularly for activity recognition from very noise input data, such as from wearable sensors [19], [20]. While most individual activity recognition approaches attempt to recognize the action of a single individual, some methods have been adapted to analyzing multiple people at once, while treating them as individuals. These include the use of hierarchical models [21] and LSTMs [22] to analyze the activities of individuals as they move.

Overall, the key limitation of individual activity recognition methods lies in that they are not able to consider a combined activity. That is, they consider individuals in isolation and fail to merge the contexts into a larger and unified intent.

B. Group Activity Recognition

Group activity recognition extends the problem of recognizing activities from multiple individuals and combines them into a single group activity. Methods have approached this from either learning the group behavior from observed single person activities, or from learning group activities from observed features in the scene.

Most commonly, approaches use a layered methodology, separating individual activities and the overall group activity. LSTMs that identified individual activities have been combined in a second layer through additional LSTMs [6] or momentum-based methods [23]. Hand crafted features based on Kalman filters were combined through layers of random forests and Markov fields [24], [25]. Layers have also been expressed as graphs, representing humans and the connections between them as separate graphs and combining them through neural networks [26], [27].

Recent approaches have utilized end-to-end neural networks, that process individual frames or videos and output an overall group activity. These have included layers of RNNs that utilized a graph-like node and edge structure [28], sets of RNNs that identify key individuals in a scene [29], and convolutional neural networks to learn interactions between individual activities and identify the group activity from these [30].

Like individual activity recognition approaches, current group activity approaches are not sufficient to identify team behaviors. While they analyze multiple individuals and consider the connections between them, they are limited to identifying behaviors where individuals are all performing the same collective activity. In real-world situations, such as those that occur in search and rescue missions, an overall team intent such as patient recovery will involve individuals performing disparate individual tasks such as communication, patient movement, and treatment. A team activity recognition approach must be able to analyze these separate activities to infer the overall intent.

III. The Proposed Approach

We will introduce the notations used in this paper. The ℓ_p-norm $(p > 0)$ of a vector $\mathbf{v}$ is defined as $\|\mathbf{v}\|_p = (\sum_i v_i^p)^{\frac{1}{p}}$. For a matrix $\mathbf{M} = [m_{ij}]$, the trace of $\mathbf{M}$ is defined as $\mathbf{tr}(\mathbf{M}) = \sum_i m_{ii}$. The $\ell_{r,p}$-norm of $\mathbf{M}$ is defined as $\|\mathbf{M}\|_{r,p} =$

$\left(\sum_{i=1}^{n}\left(\sum_{j=1}^{m}|m_{ij}|^r\right)^{\frac{p}{r}}\right)^{\frac{1}{p}} = \left(\sum_{i=1}^{n}\left\|\mathbf{m}^i\right\|_r^p\right)^{\frac{1}{p}}$, where $\mathbf{m}^i$ is the i-th column vector of $\mathbf{M}$. The Frobenius norm of $\mathbf{M}$ is defined as $\|\mathbf{M}\|_F = \sqrt{\sum_{i=1}^{n}\sum_{j=1}^{m}|m_{i,j}|^2}$. The Schatten p-norm of $\mathbf{M}$ is defined as $\|\mathbf{M}\|_{S_p} = (\sum_{i=1}^{min\{n,m\}}\sigma_i^p)^{\frac{1}{p}}$, where σ_i is the i-th singular value of $\mathbf{M}$.

Given m members within a team, the observations of the i-th team is represented as: $\mathbf{X}_i = [\mathbf{x}_i^1, \cdots, \mathbf{x}_i^m] \in \Re^{d\times m}$, where $\mathbf{x}_i^j \in \Re^d$ represents the feature vector of the observation from the j-th individual from the i-th team. We represent the team activity label vector of the observation $\mathbf{X}_i$ as $\mathbf{y}_i \in \Re^{c_1}$: if $\mathbf{X}_i$ belongs to the l-th team intent category, the l-th element within the intent label vector $\mathbf{y}_i$ satisfies $\mathbf{y}_i(l) = 1$; otherwise $\mathbf{y}_i(l) = 0$, where c_1 is the number of team intents.

Since team intent is an abstract concept that is collaboratively reflected by the activities of all individual members in the team, to model the behavioral hierarchy of the team, we introduce a latent variable $\mathbf{Z}_i = [\mathbf{z}_i^1, \cdots, \mathbf{z}_i^m] \in \Re^{c_2\times m}$ to represent individual activities from i-th team, where $\mathbf{z}_i^j \in \Re^{c_2}$ denotes the individual activity of the j-th team member from i-th team. If $\mathbf{x}_i^j$ belongs to the l-th individual activity category, the l-th element of the vector $\mathbf{z}_i^j$ satisfies $\mathbf{z}_i^j(l) = 1$; otherwise $\mathbf{z}_i^j(l) = 0$, where c_2 is the number of individual activities. This latent activity vector enable us to model concurrent individual activities, and to model diverse activities for different individuals.

Then, given a collection of n training data instances $\{(\mathbf{X}_i, \mathbf{y}_i)\}_{i=1}^n$, we define the individual activity matrix for each data instance i as $\mathbf{Z}_i$. Thus, we formulate robot recognition of team activity by minimizing the following loss function:

$$\begin{aligned}\min_{\mathbf{W},\mathbf{U},\mathbf{Z}_i,\mathbf{b},\mathbf{p}} &\sum_{i=1}^{n}(\left\|\mathbf{W}^\top\mathbf{X}_i + \mathbf{b}\mathbf{1}_m^\top - \mathbf{Z}_i\right\|_F^2 \\ &+ \left\|\mathbf{U}^\top\mathbf{Z}_i + \mathbf{p}\mathbf{1}_m^\top - \mathbf{y}_i\mathbf{1}_m^\top\right\|_F^2), \end{aligned} \tag{1}$$

where $\mathbf{U} = [\mathbf{u}_1, \cdots, \mathbf{u}_{c_1}] \in \Re^{c_2\times c_1}$ and $\mathbf{W} = [\mathbf{w}_1, \cdots, \mathbf{w}_{c_2}] \in \Re^{d\times c_2}$ represent the coefficient matrices as the parameters to be learned for team and individual activity estimation respectively. $\mathbf{b} \in \Re^{c_2\times 1}$ and $\mathbf{p} \in \Re^{c_1\times 1}$ are intercept vectors, and $\mathbf{1}_m \in \Re^{m\times 1}$ is the constant vector consisting of all 1's. The first term in Eq. (1) denotes the loss function that is designed to project from the observation data instance $\mathbf{X}_i$ to the individual activity categories $\mathbf{Z}_i$ for each individual. The second term in Eq. (1) denotes the loss function that is designed to project from latent individual activity categories $\mathbf{Z}_i$ $(i = 1, \ldots, m)$ to the team intent $\mathbf{y}_i$.

Apart from exploration of individual activities separately with latent variable vector $\mathbf{z}_i$, we further propose to uncover the relationship of individual activities among team members. Firstly, to capture the global correlations among different team members, we impose Schatten p-norm regularization to discover the common goal shared by all team members. Secondly, to preserve the local relations among teammates, we keep the local pairwise patterns in the latent subspace. To achieve this, Laplacian embedding is the right tool to leverage [31]. Specifically, we first construct a similarity matrix $\mathbf{S}_i \in \Re^{m\times m}$. Each element of $\mathbf{S}_i$ is denoted with $\mathbf{S}_i(j,k)$, where $\mathbf{S}_i(j,k)$ measures the Euclidean distance of HoG features between j-th individual $\mathbf{x}_i^j$ and k-th individual $\mathbf{x}_i^j$ from i-th team. Laplacian embedding preserves the local relationships and maximizes the smoothness of the manifold of the data in the embedding space by minimizing $\sum_{j,k=1,\cdots,m}\mathbf{S}_i(j,k)\left\|\mathbf{z}_i^j - \mathbf{z}_i^k\right\|_2^2$. Combined with teammate relationship modeling, we develop our objective function as:

$$\begin{aligned}\min_{\mathbf{W},\mathbf{U},\mathbf{Z}_i,\mathbf{b},\mathbf{p},\mathbf{Z}_i^\top\mathbf{Z}_i=\mathbf{I}} &\sum_{i=1}^{n}(\left\|\mathbf{W}^\top\mathbf{X}_i + \mathbf{b}\mathbf{1}_m^\top - \mathbf{Z}_i\right\|_F^2 \\ &+ \left\|\mathbf{U}^\top\mathbf{Z}_i + \mathbf{p}\mathbf{1}_m^\top - \mathbf{y}_i\mathbf{1}_m^\top\right\|_F^2) \\ &+ \gamma_1\sum_{i=1}^{n}\left(\|\mathbf{Z}_i\|_{S_p}^p + \sum_{j,k=1,\cdots,m}\mathbf{S}_i(j,k)\left\|\mathbf{z}_i^j - \mathbf{z}_i^k\right\|_2^2\right) \\ &+ \gamma_2\|\mathbf{W}\|_F^2 + \gamma_3\|\mathbf{U}\|_F^2, \end{aligned} \tag{2}$$

where γ_l $(l = 1, 2, 3)$ denotes trade-off hyperparameters, and $\|\mathbf{W}\|_F^2$, $\|\mathbf{U}\|_F^2$ are leveraged to avoid overfitting.

While the objective in Eq. (2) nicely defined team intent inference problem, it uses the squared Frobenius norm that is notoriously known to be very sensitive to outliers in the dataset, which may result in inferior learning performance. To improve the robustness of our model [8], [9], [10], [11], we substitute squared ℓ_2-norm with $\ell_{2,p}$-norm and $\ell_{p,p}$-norm $(0 < p \leq 2)$ as follows:

$$\begin{aligned}\min_{\mathbf{W},\mathbf{U},\mathbf{Z}_i,\mathbf{b},\mathbf{p},\mathbf{Z}_i^\top\mathbf{Z}_i=\mathbf{I}} &\sum_{i=1}^{n}(\left\|\mathbf{W}^\top\mathbf{X}_i + \mathbf{b}\mathbf{1}_m^\top - \mathbf{Z}_i\right\|_{2,p}^p \\ &+ \left\|\mathbf{U}^\top\mathbf{Z}_i + \mathbf{p}\mathbf{1}_m^\top - \mathbf{y}_i\mathbf{1}_m^\top\right\|_{2,p}^p) \\ &+ \gamma_1\sum_{i=1}^{n}\left(\|\mathbf{Z}_i\|_{S_p}^p + \sum_{j,k=1,\cdots,m}\mathbf{S}_i(j,k)\left\|\mathbf{z}_i^j - \mathbf{z}_i^k\right\|_2^p\right) \\ &+ \gamma_2\|\mathbf{W}\|_F^2 + \gamma_3\|\mathbf{U}\|_F^2, \end{aligned} \tag{3}$$

Upon solving the regularized optimization problem in Eq. (3) and obtaining the optimal parameters, we can use the learned model for robots to recognize the team intent in an online fashion. The team intent is computed as follows:

$$\textit{Team intent} \;= \arg\max \mathbf{y}_o(l),\; l = 1, 2, \cdots, c_1, \tag{4}$$

where $\mathbf{y}_o(l) = \frac{1}{m}\mathbf{U}^\top\mathbf{Z}_o\mathbf{1}_m + \mathbf{p}$ and $\mathbf{Z}_o = \mathbf{W}^\top\mathbf{X}_o + \mathbf{b}\mathbf{1}_m^\top$, $\mathbf{Z}$ and $\mathbf{U}$ denote the optimal coefficient matrices learned in the training process and $\mathbf{X}_o$ is the query observation.

IV. Optimization Algorithm

Although the motivation of the formulation of our new method in Eq. (3) is clear and justifiable, it is a non-smooth objective, which is difficult to efficiently solve in general. Motivated by our earlier works that use the iterative reweighted method [8], [10], [32] to solve non-smooth objectives and

taking into account of the issues of its numerical stability [11], we rewrite the objective in Eq. (3) as follows:

$$\begin{aligned}
&\min_{\mathbf{W},\mathbf{U},\mathbf{Z}_i,\mathbf{b},\mathbf{p},\mathbf{Z}_i^\top\mathbf{Z}_i=\mathbf{I}} \sum_{i=1}^{n} \mathrm{tr}((\mathbf{W}^\top\mathbf{X}_i+\mathbf{b}\mathbf{1}_m^\top-\mathbf{Z}_i)^\top \tilde{\mathbf{D}}_i \\
&(\mathbf{W}^\top\mathbf{X}_i+\mathbf{b}\mathbf{1}_m^\top-\mathbf{Z}_i)) + \sum_{i=1}^{n} \mathrm{tr}((\mathbf{U}^\top\mathbf{Z}_i+\mathbf{p}\mathbf{1}_m^\top-\mathbf{y}_i\mathbf{1}_m^\top)^\top \\
&\hat{\mathbf{D}}_i(\mathbf{U}^\top\mathbf{Z}_i+\mathbf{p}\mathbf{1}_m^\top-\mathbf{y}_i\mathbf{1}_m^\top)) + \gamma_1 \sum_{i=1}^{n} (\mathrm{tr}(\mathbf{Z}_i^\top\bar{\mathbf{D}}_i\mathbf{Z}_i) \\
&+ \sum_{j,k=1,\cdots,m} \mathbf{S}_i(j,k)\mathbf{\Theta}_i(j,k)\left\|\mathbf{z}_i^j-\mathbf{z}_i^k\right\|_2^2) + \gamma_2\|\mathbf{W}\|_F^2 \\
&+ \gamma_3\|\mathbf{U}\|_F^2,
\end{aligned} \tag{5}$$

where $\tilde{\mathbf{D}}_i$ is a diagonal matrix whose k-th element is $\frac{p}{2}\left(\left\|\tilde{\mathbf{e}}_i^k\right\|_2^2+\delta\right)^{\frac{p-2}{2}}$, $\tilde{\mathbf{e}}_i^k$ is k-th column vector of $\tilde{\mathbf{E}}_i = \mathbf{W}^\top\mathbf{X}_i+\mathbf{b}\mathbf{1}_m^\top-\mathbf{Z}_i$. $\hat{\mathbf{D}}_i$ is also a diagonal matrix whose k-th element $\frac{p}{2}\left(\left\|\hat{\mathbf{e}}_i^k\right\|_2^2+\delta\right)^{\frac{p-2}{2}}$, $\hat{\mathbf{e}}_i^k$ is k-th column vector of $\hat{\mathbf{E}}_i = \mathbf{U}^\top\mathbf{Z}_i+\mathbf{p}\mathbf{1}_m^\top-\mathbf{y}_i\mathbf{1}_m^\top$. $\bar{\mathbf{D}}_i = \frac{p}{2}(\mathbf{Z}_i^\top\mathbf{Z}_i+\sigma\mathbf{I})^{\frac{p-2}{2}}$ and $\mathbf{\Theta}_i(j,k) = \frac{p}{2}\left(\left\|\mathbf{z}_i^j-\mathbf{z}_i^k\right\|_2^2+\delta\right)^{\frac{p-2}{2}}$.

Before giving the solution algorithm to optimize Eq. (5), we will first introduce the Alternating Direction Method of Multipliers (ADMM), which was proposed in [33], [34] to solve convex optimization problems by breaking them into smaller pieces that are easier to handle.

Specifically, given the following objective with the equality constraint:

$$\min_{x,z} f(x)+g(z), \qquad s.t. \quad h(x,z)=0, \tag{6}$$

Algorithm 1 solves the problem by decoupling it into subproblems and optimizing each variable while fixing others [33], [34], where y is the Lagrangian multiplier to the constraint h. It is worth noting that Algorithm 1 was proved to converge Q-linearly to the optimal solution [33].

Following the framework of ADMM, we further rewrite the objective in Eq. (5) as follows:

$$\begin{aligned}
&\min_{\mathbf{W},\mathbf{U},\mathbf{Z}_i,\mathbf{b},\mathbf{p},} \sum_{i=1}^{n} \mathrm{tr}((\mathbf{W}^\top\mathbf{X}_i+\mathbf{b}\mathbf{1}_m^\top-\mathbf{Z}_i)^\top \tilde{\mathbf{D}}_i \\
&(\mathbf{W}^\top\mathbf{X}_i+\mathbf{b}\mathbf{1}_m^\top-\mathbf{Z}_i)) + \sum_{i=1}^{n} \mathrm{tr}((\mathbf{U}^\top\mathbf{Z}_i+\mathbf{p}\mathbf{1}_m^\top-\mathbf{y}_i\mathbf{1}_m^\top)^\top \\
&\hat{\mathbf{D}}_i(\mathbf{U}^\top\mathbf{Z}_i+\mathbf{p}\mathbf{1}_m^\top-\mathbf{y}_i\mathbf{1}_m^\top)) + \gamma_1 \sum_{i=1}^{n} (\mathrm{tr}(\mathbf{Z}_i^\top\bar{\mathbf{D}}_i\mathbf{E}_i) \\
&+ \mathrm{tr}(\mathbf{Z}_i^\top\bar{\mathbf{L}}_i\mathbf{E}_i)) + \gamma_2\|\mathbf{W}\|_F^2 + \gamma_3\|\mathbf{U}\|_F^2 \\
&+ \sum_{i=1}^{n} \frac{\mu}{2}\left\|\mathbf{Z}_i-\mathbf{E}_i+\frac{\mathbf{\Lambda}_i}{\mu}\right\|_F^2, \quad s.t. \quad \mathbf{E}_i^\top\mathbf{E}_i=\mathbf{I},
\end{aligned} \tag{7}$$

where $\tilde{\mathbf{S}}_i \in \Re^{m\times m}$ is the reconstructed similarity matrix whose element value $\tilde{\mathbf{S}}_i(j,k) = \Theta_i(j,k)\mathbf{S}_i(j,k)$ and $\mathbf{L}_i = \mathbf{D}_i - \tilde{\mathbf{S}}_i$ where $\mathbf{D}_i$ is a diagonal matrix whose entries are column (or row) sum of $\tilde{\mathbf{S}}_i$. The j-th diagonal element of $\mathbf{D}_i$ is $\sum_j \tilde{\mathbf{S}}_i(j,k)$. $\mathbf{\Lambda}_i$ is the Lagrangian multiplier for the constraint of $\mathbf{Z}_i = \mathbf{E}_i$. For brevity, we define $\mathbf{F}_i = \mathbf{W}^\top\mathbf{X}_i+\mathbf{b}\mathbf{1}_m^\top$ and $\mathbf{P}_i = \mathbf{y}_i\mathbf{1}_m^\top-\mathbf{p}\mathbf{1}_m^\top$. The solution for Eq. (7) is summarized in Algorithm 2.

Algorithm 1: The ADMM algorithm.

Set $1<\rho<2$ and initialize $\mu>0$ and y;
while *not converge* **do**
1. Update x by solving $x^{k+1} = \arg\min_x(f(x)+\frac{\mu}{2}\|h(x,z^k)+\frac{y^k}{\mu}\|^2)$;
2. Update z by solving $z^{k+1} = \arg\min_z(g(z)+\frac{\mu}{2}\|h(x^{k+1},z)+\frac{y^k}{\mu}\|^2)$;
3. Update y by $y^{k+1} = y^k+\mu h(x^{k+1},z^{k+1})$;
4. Update μ by $\mu=\rho\mu$.

end

Algorithm 2: Solve the optimization problem in Eq. (7).

Initialization $\mathbf{W}$, $\mathbf{b}$, $\mathbf{U}$, $\mathbf{p}$, $\mathbf{Z}_i$, $\mathbf{E}_i$, $\mathbf{\Lambda}_i$, $0<\rho<2$, $\mu,\gamma_1,\gamma_2,\gamma_3>0$;
while *not converge* **do**
1. Update $\mathbf{b}$ by $\mathbf{b} = (\sum_{i=1}^n m\tilde{\mathbf{D}}_i)^{-1}(\sum_{i=1}^n \tilde{\mathbf{D}}_i(\mathbf{Z}_i-\mathbf{W}^\top\mathbf{X}_i)\mathbf{1}_m)$;
2. Update $\mathbf{W}$ by $\mathbf{W} = (\sum_{i=1}^n p\tilde{\mathbf{D}}_i\mathbf{X}_i\mathbf{X}_i^\top+\gamma_2\mathbf{I}_d)^{-1}(\sum_{i=1}^n \tilde{\mathbf{D}}_i\mathbf{X}_i(\mathbf{b}\mathbf{1}_m^\top-\mathbf{Z}_i))$;
3. Update $\mathbf{p}$ by $\mathbf{p} = (\sum_{i=1}^n m\hat{\mathbf{D}}_i)^{-1}(\sum_{i=1}^n \hat{\mathbf{D}}_i(\mathbf{y}_i\mathbf{1}_m^\top-\mathbf{U}^\top\mathbf{Z}_i)\mathbf{1}_m)$;
4. Update $\mathbf{U}$ by $\mathbf{U} = (\sum_{i=1}^n \hat{\mathbf{D}}_i\mathbf{Z}_i\mathbf{Z}_i^\top+\gamma_3\mathbf{I}_{c_2})^{-1}(\sum_{i=1}^n \hat{\mathbf{D}}_i\mathbf{Z}_i(\mathbf{p}\mathbf{1}_m^\top-\mathbf{y}_i\mathbf{1}_m^\top))$;
5. Update $\mathbf{Z}_i$ by $\mathbf{Z}_i = (2\tilde{\mathbf{D}}_i+2\mathbf{U}\hat{\mathbf{D}}_i\mathbf{U}^\top+\mu\mathbf{I})^{-1}(2\tilde{\mathbf{D}}_i\mathbf{F}_i+2\mathbf{U}\hat{\mathbf{D}}_i\mathbf{P}_i+\gamma_1\bar{\mathbf{D}}_i\mathbf{E}_i+\gamma_1\mathbf{L}_i\mathbf{E}_i+\mu\mathbf{E}_i-\mathbf{\Lambda}_i)$;
6. Update $\mathbf{E}_i$ by $\mathbf{E}_i = \mathbf{U}\mathbf{V}^\top$ where $\mathbf{M}_i = \mathbf{Z}_i-\mathbf{L}_i\mathbf{Z}_i-\bar{\mathbf{D}}_i\mathbf{Z}_i+\frac{\mathbf{\Lambda}_i}{\mu}$ and $\mathrm{svd}(\mathbf{M}_i) = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^T$;
7. Update $\mathbf{\Lambda}_i$ by $\mathbf{\Lambda}_i = \mathbf{\Lambda}_i+\mu(\mathbf{Z}_i-\mathbf{E}_i)$;
8. Update μ by $\mu=\rho\mu$;

end
Output: $\mathbf{W}$, $\mathbf{b}$, $\mathbf{U}$, $\mathbf{p}$.

The general ADMM method has been proved to converge to the optimal solution [33]. Given iterations $k=0,1,\ldots,K$, convergence requires that $0<\mu_k<\mu_{k+1}$ for all k, $\mu_k\to\infty$. Under this assumption, the current solution $\mathbf{X}_k$ will approach the optimal solution $\mathbf{X}^*$. Since Algorithm 2 defines that $0<\rho<2$ and updates μ by $\mu=\rho\mu$, this condition always holds for our solution.

Additionally, the computational complexity of the algorithm derived using the ADMM depends on the objective function $f(\mathbf{X})$, which is $\mathcal{J}(\mathbf{Z})$ defined in Eq. (7). The complexity of $\mathcal{J}(\mathbf{Z})$ thereby is

$$\mathcal{O}(i(mc_2(d+1+c_1+c_2+m)+mc_1+m)), \tag{8}$$

where i is the number of training instances, m is the number of team members, c_2 is the number of individual activities, c_1 is the number of team intents, and d is the dimensionality of the representation for each team member. This complexity is linear with respect to any individual model parameter. For recognizing team intents with a trained model, the formulation is again linear with respect to any single model parameter. The

complexity of a single team intent recognition thereby is

$$\mathcal{O}(mc_2(d + c_1)). \tag{9}$$

V. EXPERIMENTS

In this section, we empirically evaluate the performance of our proposed approach over team behaviors recognition using two benchmark datasets. In addition, ablation study is conducted to validate our approach, executing our approach on the limited computing power available on that platform. These experiments demonstrate that our approach can perform well in a challenging teaming scenarios, compares favorably with the current state-of-the-art, and can be used in real-time on a mobile robot platform.

A. Collective Activity Dataset

The Collective Activity Dataset (CAD) [5] is a benchmark group activity dataset used in the computer vision community to evaluate group activity recognition approaches. CAD consists of videos of varying size, resolution, and length of variously sized groups performing different activities (crossing, talking, dancing and jogging). The group activity label for a frame is defined by the activity in which most people participant. For each individual in the team, we extract histogram of oriented gradients (HoG) [35] from each modality (rgb, depth, and thermal) for each actor with provided ground truth bounding boxes. To evaluate our approach on CAD, we conduct a 5-fold cross-validation approach and compute accuracy of our prediction of the group activity. We iterate each five-fold experiment 10 times and randomly shuffle training and testing groups in between each iteration. The average performance for a given model with fixed hyperparameters are used for comparison. The hyperparameters γ_l $(l = 1, 2, 3)$ in our model are fined tuned by a search on $\{10^{-4}, 10^{-3}, \ldots, 10^{4}\}$ and hyperparameters p are fined tuned by a search on $\{0.3, 0.5, 0.8, 1.0, 1.5, 2.0\}$.

Table I compares results of our new method against other state-of-the-art group activity recognition methods on the dataset, including: (1) VGG-16 [36], a deep convolutional neural network; (2) LRCN [37], Long-term Recurrent Convolutional Networks; (3) VGG-16-Person, a deep convolutional neural network for person recognition and (4) LRCN-Person [23], a Long-term Recurrent Convolutional Networks for person recognition. These are all approaches based on large deep learning networks. We also compared to two baseline machine learning methods that also run in real-time, a multiclass Gaussian kernel Support Vector Machine (SVM), a nearest neighbor approach and a logistic regression model.

In terms of accuracy, we can see that proposed hierarchical team behavior recognition approach achieves a better performance comparing the traditional machine learning models SVM and nearest neighbors. It can be attributed to the following reasons. Due to the introduction of latent variable $\mathbf{Z}$, our hierarchical model could capture the team member activities, which is beneficial for team intent recognition. Moreover, via the leverage of Schatten p-norm and Laplacian embedding, we could explore the teammates structure globally and locally. The interacting among team members are obtained to facilitate the recognition of team intent. Our approach performs comparably well to the state-of-the-art complex deep learning based approaches. In addition, our model is interpretable compare to learning based approaches. The weights learned from our model indicates the importance of features or individuals in our model. The regularization terms are also leveraged to discover the relationship between individuals. Moreover, in our hierarchy team behaviour recognition, value of p is an important hyperparameter. When the value of p is high, our model will be sensitive to the outlier and these few outliers dominate our model, which lead in a recognition accuracy drop. When the value of p is low, the training sample will contribute equally to our model which also degrade for our team behaviour recognition performance. From our extensive experiments our approach achieves its peak performance when the hyperparameter p is set to be 0.8.

TABLE I
RESULTS OF ACCURACY AND REAL-TIME PERFORMANCE ON THE CAD DATASET, AND COMPARISONS WITH OTHER STATE-OF-THE-ART AND BASELINE METHODS.

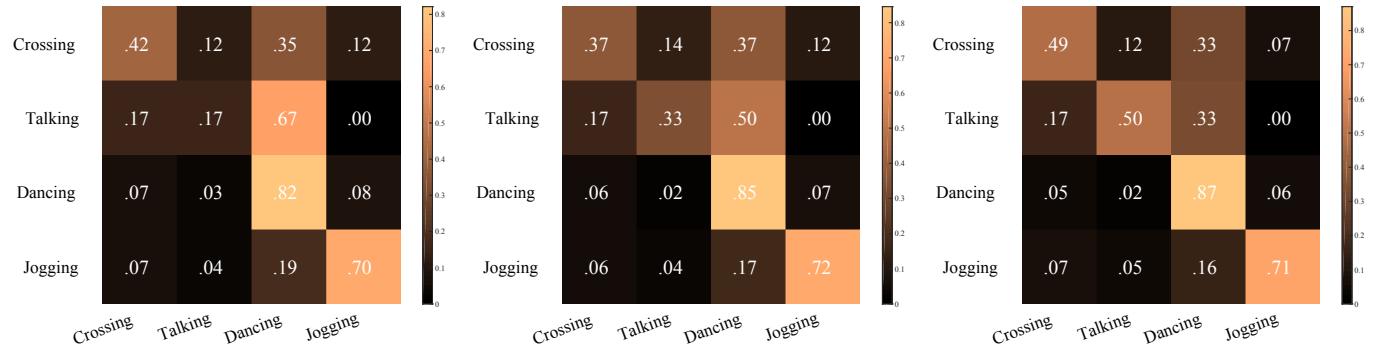
Method		Accuracy	Real-Time
SVM		65.87%	Y
Nearest Neighbors		38.52%	Y
Logistic Regression		61.37%	Y
VGG-16 ([36])		68.3%	N
LRCN ([37])		64.2%	N
VGG-16-Person ([36])		71.2%	N
LRCN-Person ([23])		64.0%	N
Our Approach (Without $\mathbf{Z}$)	p = 2.0	57.13%	Y
	p = 1.5	53.04%	Y
	p = 1.0	58.21%	Y
	p = 0.8	56.15%	Y
	p = 0.5	54.81%	Y
	p = 0.3	49.32%	Y
Our Approach (With $\mathbf{Z}$)	p = 2.0	73.41%	Y
	p = 1.5	73.02%	Y
	p = 1.0	71.85%	Y
	p = 0.8	74.60%	Y
	p = 0.5	71.43%	Y
	p = 0.3	69.84%	Y

In terms of real-time performance, we consider a processing speed of more than ten frames as *real-time*, which is similar to the processing speed of a human visual system. With fixed hyper parameters p, it takes 209.94s for training on 500 CAD samples, ending in 175 iterations, and takes 2.26s for inference of 257 CAD samples. The existing deep learning approaches require days for training and have poor run time on board robots, which make them unsuitable for robotics applications.

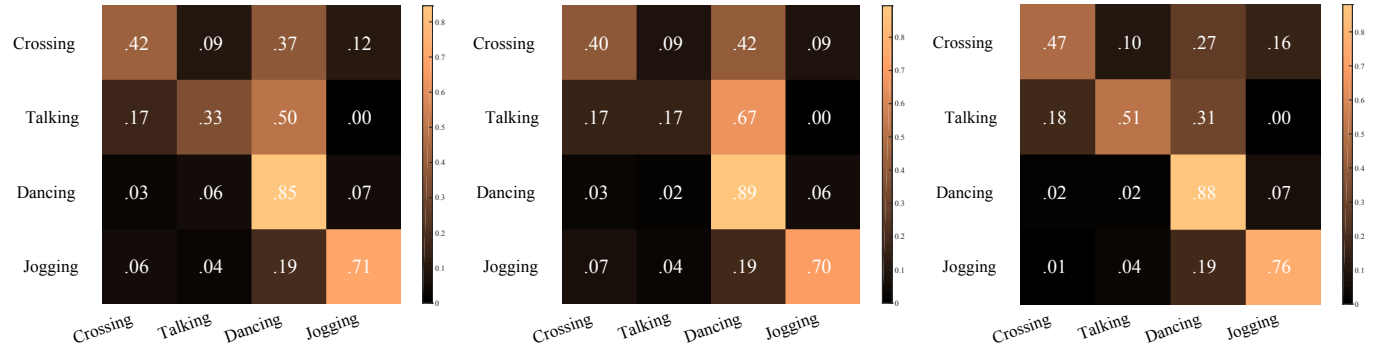
Besides the overall the prediction accuracy reported in Table I, The accuracy by behavior category of our approach under different value p is illustrated in Fig. 2. From Fig. 2, we can see that our model achieves a stable predictive capability by behavior category under different p.

B. Effect of Teammate Relationship Term

Apart from the performance compassion between our approach with the state of art, we also study the degenerate



(a) p=0.3 (b) p=0.5 (c) p=0.8



(d) p=1.0 (e) p=1.5 (f) p=2.0

Fig. 2. Confusion matrix illustrating accuracy by behavior category of our approach (under different p) on the Collective Activity Dataset.

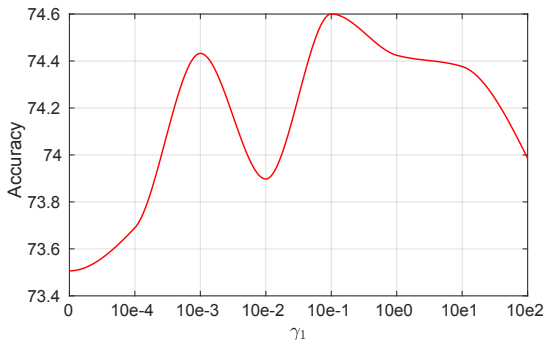


Fig. 3. Graph showing the effect of γ_1, the hyperparameter controlling the effect of teammate relationship modeling

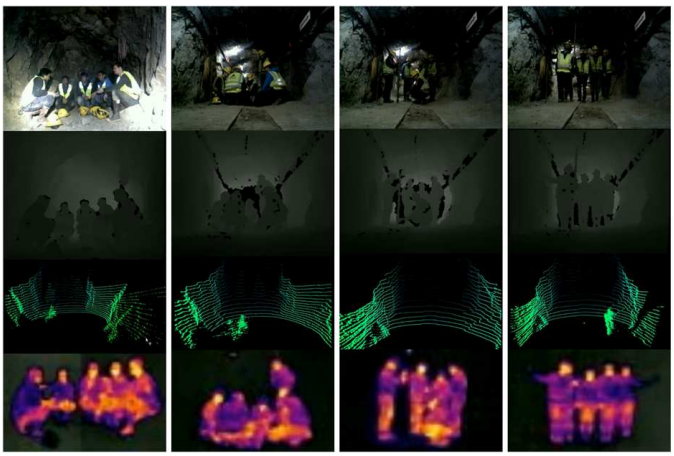


Fig. 4. Example data instances from our own dataset, consisting of multimodal perception data captured in the dark side of the Edgar Mine by RGB-D camera, LiDAR, and thermal camera simultaneously. Each column shows different team intents. Left to right: *donning*, *patient care*, *timbering*, and *traversing*.

version of our approach without leveraging the latent variable term $\mathbf{Z}$, whose loss function can be rewritten as follows:

$$\min_{\mathbf{W},\mathbf{U},\mathbf{b}} \sum_{i=1}^{n} \left\| \mathbf{W}^\top \mathbf{X}_i + \mathbf{b}\mathbf{1}_m^\top - \mathbf{y}_i \mathbf{1}_m^\top \right\|_{2,p}^p + \gamma_1 \|\mathbf{W}\|_F^2. \quad (10)$$

The team intent is computed as follows:

$$\textit{Team intent} \ = \arg\max \mathbf{y}_o(l), \ l = 1, 2, \cdots, c_1, \quad (11)$$

where $\mathbf{y}_o(l) = \frac{1}{m}\mathbf{W}^\top \mathbf{X}_o \mathbf{1}_m + \mathbf{b}$, $\mathbf{W}$ and $\mathbf{b}$ denote the optimal coefficient matrices learned in the training process and $\mathbf{X}_o$ is the query observation. From Table I, we can see that our approach consistently achieves a better performance compared to its degenerate version, which demonstrates the effectiveness of latent variables $\mathbf{Z}$ in our model.

In addition, we explore the effect of teammate relationship term utilized in our model. Fig. 3 illustrates the performance accuracy w.r.t to γ_1, the hyperparameter weighting the $\mathbf{Z}$ regularization that uncovers the teammate correlations. This graph demonstrates the importance of γ_1 to the team intent recognition task in CAD. It demonstrates that, with a fixed $\gamma_2 = 10^{-3}$, $\gamma_3 = 10^{-2}$ and $p = 0.8$, intent recognition accuracy is at its lowest point when $\gamma_1 = 0$, when teammates correlation regularization is not used. As the value of γ_1 increases along the log scale with $\gamma_1 = 10^{-2}$, our approach achieves higher accuracy, peaking at 74.60%.

TABLE II
RESULTS ON THE OUR OWN DATASET. FOR THIS DATASET, WE COMPARED ONLY TO REAL-TIME METHODS.

Method		Accuracy
SVM		75.29%
Nearest Neighbors		81.18%
Logistic Regression		89.42%
Our Approach	p = 2.0	95.65%
	p = 1.5	96.47%
	p = 1.0	95.29%
	p = 0.8	94.12%
	p = 0.5	94.12%
	p = 0.3	94.12%

C. Real-World Search and Rescue Dataset

We then evaluate our approach of prediction of team intent on the multisensory underground search and rescue teamwork dataset, which is collected at our own lab. This team behavior dataset is unique in that it is collected underground, with team intents chosen to correspond to real-world search and rescue tasks. The five team intents defined are *donning*, *patient care*, *team stop*, *timbering* and *Traversing*. The team was recorded by a Husky robot [1] equipped with an RGB-D camera, a thermal camera, and LiDAR, simultaneously resulting in color images, depth images, point clouds, and thermal images. Fig. 4 illustrates this multisensory data as the team performed *donning*, *patient care*, *timbering*, and *traversing*, respectively. Each team behavior was performed 20 times in different team configurations (i.e., roles in the team were reconfigured for each execution, in order to record instances where different body scales and motion patterns would be used). Ground truth data was labeled manually. In our evaluation, we utilized the color sensory data and extracted HoG features each individual from these multi-sensor observations to create a bag-of-words representation for each team member. We conduct a 5-fold cross validation approach to obtain the optimal model parameters. Then, Eq. (4) is applied to the remaining data instances to recognize the team intent occurring in each scene.

By applying our proposed approach, we achieved an accuracy rate of 96.47%. In term of real time performance, it takes 0.92s for training on 85 real world mining samples, ending in 21 iterations, and takes 0.11s for inference of 25 real world mining samples.

In the real-world search and rescue scenarios where these missions performed are often hazardous, it is required that the robots can recognize team activities in real time. Thus, in our collected real-world dataset, we only compared to two baseline methods, both able to operate in real-time, using the same bag-of-words representation. A multi-class Gaussian kernel Support Vector Machine identified only 75.29% of the team intents correctly. A nearest neighbors based approach performed slightly better, correctly recognizing 81.18% of the team intents. A logistic regression model achieved a better performance with 89.42% recognition accuracy. From Table II, we can see that our approach achieve its best performance when $p = 1.5$ with fixed parameters $\gamma_1 = 1$, $\gamma_2 = 10^{-3}$ and $\gamma_3 = 10^{-1}$. This demonstrates the superior performance of our approach.

[1]Husky Unmanned Ground Vehicle: www.clearpathrobotics.com/husky-unmanned-ground-vehicle-robot.

VI. CONCLUSION

We propose a novel, real-time approach for recognition of team behaviors from multisensory data. We formulate team behavior recognition as a joint learning problem, learning a projection from multisensory observations to individual activities and learning a projection from individual activities to an overall team intent simultaneously in the same model. We model these learned individual activity labels as a latent variable. We also introduce an explicit representation of the relationships among members of the team. This is represented by a teammate interaction graph, which is then projected using graph embedding into a vector representation. This vector representation, encoding teammate relationships, is used to learn the latent individual activity labels. We formulate our method in a unified optimization framework, and introduce a new optimization algorithm that theoretically converges to the optimal solution.

To evaluate our approach, extensive experiments are performed. We show that our approach performs competitively on a benchmark group activity dataset, while still running in real-time. Apart from the recognition performance and running time, the effect of teammate interaction graph is studied. We then introduce the multisensory real-word search and rescue dataset, consisting of observations made by a Husky mobile robot in an underground mine, and show that our approach outperforms baseline real-time methods, and that our introduction of the graph embedded teammate interaction graph increases our methods performance. Finally, we perform an ablation study of our method to explore the effects of hyperparameters. These results show that our approach achieves superior accuracy and real-time performance, making it the ideal method for team intent recognition on mobile robot platforms.

REFERENCES

[1] M. Erdelj, E. Natalizio, K. R. Chowdhury, and I. F. Akyildiz, "Help from the sky: Leveraging uavs for disaster management," *IEEE Pervasive Computing*, vol. 16, no. 1, pp. 24–32, 2017.
[2] R. Schulz, P. Kratzer, and M. Toussaint, "Preferred interaction styles for human-robot collaboration vary over tasks with different action types," *Frontiers in neurorobotics*, vol. 12, p. 36, 2018.
[3] F. Han, B. Reily, W. Hoff, and H. Zhang, "Space-time representation of people based on 3d skeletal data: A review," *Computer Vision and Image Understanding*, vol. 158, pp. 85–105, 2017.
[4] V. Ramanathan, B. Yao, and L. Fei-Fei, "Social role discovery in human events," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2013, pp. 2475–2482.
[5] W. Choi, K. Shahid, and S. Savarese, "What are they doing?: Collective activity classification using spatio-temporal relationship among people," in *2009 IEEE 12th International Conference on Computer Vision Workshops, ICCV Workshops*. IEEE, 2009, pp. 1282–1289.
[6] M. S. Ibrahim, S. Muralidharan, Z. Deng, A. Vahdat, and G. Mori, "A hierarchical deep temporal model for group activity recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 1971–1980.
[7] M. Qi, J. Qin, A. Li, Y. Wang, J. Luo, and L. Van Gool, "stagnet: An attentive semantic rnn for group activity recognition," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 101–117.

[8] F. Nie, H. Wang, X. Cai, H. Huang, and C. Ding, "Robust matrix completion via joint schatten p-norm and ℓ_p-norm minimization," in *2012 IEEE 12th International Conference on Data Mining*. IEEE, 2012, pp. 566–574.

[9] H. Wang, F. Nie, W. Cai, and H. Huang, "Semi-supervised robust dictionary learning via efficient $\ell_{2,0^+}$-norms minimization," in *Proceedings of the IEEE International Conference on Computer Vision*, 2013, pp. 1145–1152.

[10] F. Nie, H. Wang, H. Huang, and C. Ding, "Joint schatten p-norm and ℓ_p-norm robust matrix completion for missing value recovery," *Knowledge and Information Systems*, vol. 42, no. 3, pp. 525–544, 2015.

[11] H. Yang, K. Liu, H. Wang, and F. Nie, "Learning strictly orthogonal p-order nonnegative laplacian embedding via smoothed iterative reweighted method," in *Proceedings of the 28th International Joint Conference on Artificial Intelligence*. AAAI Press, 2019, pp. 4040–4046.

[12] J. Heard, R. Heald, C. E. Harriott, and J. A. Adams, "A diagnostic human workload assessment algorithm for human-robot teams," in *Companion of the 2018 ACM/IEEE International Conference on Human-Robot Interaction*, 2018, pp. 123–124.

[13] S. al Mahi, M. Atkins, and C. Crick, "Learning to assess the cognitive capacity of human partners," in *Proceedings of the Companion of the 2017 ACM/IEEE International Conference on Human-Robot Interaction*, 2017, pp. 63–64.

[14] Y. Zhang, V. Narayanan, T. Chakraborti, and S. Kambhampati, "A human factors analysis of proactive support in human-robot teaming," in *2015 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2015, pp. 3586–3593.

[15] T. Fong, "Human-robot teaming: Communication, coordination, and collaboration," 2017.

[16] J. Wang, Z. Liu, Y. Wu, and J. Yuan, "Learning actionlet ensemble for 3d human action recognition," *IEEE transactions on pattern analysis and machine intelligence*, vol. 36, no. 5, pp. 914–927, 2013.

[17] G. Yu, Z. Liu, and J. Yuan, "Discriminative orderlet mining for real-time recognition of human-object interaction," in *Asian Conference on Computer Vision*. Springer, 2014, pp. 50–65.

[18] B. Reily, F. Han, L. E. Parker, and H. Zhang, "Skeleton-based bio-inspired human activity prediction for real-time human–robot interaction," *Autonomous Robots*, vol. 42, no. 6, pp. 1281–1298, 2018.

[19] F. Ordóñez and D. Roggen, "Deep convolutional and lstm recurrent neural networks for multimodal wearable activity recognition," *Sensors*, vol. 16, no. 1, p. 115, 2016.

[20] C. A. Ronao and S.-B. Cho, "Human activity recognition with smartphone sensors using deep learning neural networks," *Expert systems with applications*, vol. 59, pp. 235–244, 2016.

[21] M. S. Ryoo and J. K. Aggarwal, "Spatio-temporal relationship match: Video structure comparison for recognition of complex human activities," in *2009 IEEE 12th international conference on computer vision*. IEEE, 2009, pp. 1593–1600.

[22] A. Alahi, K. Goel, V. Ramanathan, A. Robicquet, L. Fei-Fei, and S. Savarese, "Social lstm: Human trajectory prediction in crowded spaces," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 961–971.

[23] T. Shu, S. Todorovic, and S.-C. Zhu, "Cern: confidence-energy recurrent network for group activity recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 5523–5531.

[24] W. Choi, K. Shahid, and S. Savarese, "Learning context for collective activity recognition," in *CVPR*, 2011, pp. 3273–3280.

[25] W. Choi and S. Savarese, "Understanding collective activitiesof people from videos," *IEEE transactions on pattern analysis and machine intelligence*, vol. 36, no. 6, pp. 1242–1257, 2013.

[26] Z. Deng, A. Vahdat, H. Hu, and G. Mori, "Structure inference machines: Recurrent neural networks for analyzing relations in group activity recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 4772–4781.

[27] W. Li, M.-C. Chang, and S. Lyu, "Who did what at where and when: Simultaneous multi-person tracking and activity recognition," *arXiv preprint arXiv:1807.01253*, 2018.

[28] S. Biswas and J. Gall, "Structural recurrent neural network (srnn) for group activity analysis," in *WACV*, 2018, pp. 1625–1632.

[29] V. Ramanathan, J. Huang, S. Abu-El-Haija, A. Gorban, K. Murphy, and L. Fei-Fei, "Detecting events and key actors in multi-person videos," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 3043–3053.

[30] Z. Deng, M. Zhai, L. Chen, Y. Liu, S. Muralidharan, M. J. Roshtkhari, and G. Mori, "Deep structured models for group activity recognition," *arXiv preprint arXiv:1506.04191*, 2015.

[31] H. Wang, F. Nie, and H. Huang, "Learning robust locality preserving projection via p-order minimization," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 29, no. 1, 2015.

[32] Y. Liu, Y. Guo, H. Wang, F. Nie, and H. Huang, "Semi-supervised classifications via elastic and robust embedding," in *AAAI*, 2017.

[33] D. P. Bertsekas, *Constrained optimization and Lagrange multiplier methods*. Athena Scientific, 1996.

[34] S. Boyd, N. Parikh, E. Chu, B. Peleato, J. Eckstein *et al.*, "Distributed optimization and statistical learning via the alternating direction method of multipliers," *Foundations and Trends® in Machine learning*, vol. 3, no. 1, pp. 1–122, 2011.

[35] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *2005 IEEE computer society conference on computer vision and pattern recognition (CVPR'05)*, vol. 1. IEEE, 2005, pp. 886–893.

[36] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[37] J. Donahue, L. Anne Hendricks, S. Guadarrama, M. Rohrbach, S. Venugopalan, K. Saenko, and T. Darrell, "Long-term recurrent convolutional networks for visual recognition and description," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 2625–2634.