Human Action Recognition via Skeletal and Depth based Feature Fusion

Meng Li*
City University of Hong Kong

Howard Leung†
City University of Hong Kong

Hubert P. H. Shum‡
Northumbria University

Abstract

This paper addresses the problem of recognizing human actions captured with depth cameras. Human action recognition is a challenging task as the articulated action data is high dimensional in both spatial and temporal domains. An effective approach to handle this complexity is to divide human body into different body parts according to human skeletal joint positions, and performs recognition based on these part-based feature descriptors. Since different types of features could share some similar hidden structures, and different actions may be well characterized by properties common to all features (sharable structure) and those specific to a feature (specific structure), we propose a joint group sparse regression-based learning method to model each action. Our method can mine the sharable and specific structures among its part-based multiple features meanwhile imposing the importance of these part-based feature structures by joint group sparse regularization, in favor of discriminative part-based feature structure selection. To represent the dynamics and appearance of the human body parts, we employ part-based multiple features extracted from skeleton and depth data respectively. Then, using the group sparse regularization techniques, we have derived an algorithm for mining the key part-based features in the proposed learning framework. The resulting features derived from the learnt weight matrices are more discriminative for multi-task classification. Through extensive experiments on three public datasets, we demonstrate that our approach outperforms existing methods.

Keywords: action recognition, regularization, feature fusion, group sparse

Concepts: •**Computing methodologies** → **Activity recognition and understanding; Motion capture;**

1 Introduction

Human action recognition has many potential applications including video games, surveillance, robotics, etc. Despite the research efforts in the past decade and many encouraging advances, it is still challenging to have accurate action recognition due to the high dimensional and articulated nature of human actions performed under a variety of scenarios. In addition, some actions may involve interactions with external objects in the environment, which increases the difficulty of action recognition. In this paper, we focus on human action recognition in depth videos that have recently driven significant attention from researchers [Han et al. 2016; Aggarwal and Xia 2014]. The 3D locations of skeletal joints provided by the skeleton estimation algorithm [Shotton et al. 2013] make it easier to represent a human motion as a set of movements of body parts.

Although skeletal features are very helpful for human action recognition, they may not work well on certain occasions because: (1) the 3D positions of the tracked joints in the depth video are not always accurate, which increases the intra-class variations in the actions, and (2) it is insufficient to use only the 3D joint positions to fully model a human action, especially when the action includes the interactions between human and objects.

To alleviate these problems, different appearance features based on the depth data can be leveraged. [Wang et al. 2014] proposed the Local Occupancy Patterns (LOP) as the local depth appearance for each joint to characterize the interaction between the human subject and the objects. Histogram of Oriented Principal Component (HOPC) [Rahmani et al. 2014a], which is another local depth appearance feature for each joint, gives more informative and robust model around the joints. As different features may perform optimally under different conditions, it is reasonable to combine these features so that they complement each other. Such multiple features may contain some common properties among all feature sets (i.e. sharable structures) while each feature set may possess its own unique characteristics (i.e. specific structures). As a result, it is important to extract the sharable and specific structures from the multiple features for multi-task classification [Chen et al. 2013; Zhang and Yeung 2012; Amit et al. 2007; Amit et al. 2007], which can significantly reduce the complexity of the task due to share information between related tasks [Amit et al. 2007; Torralba et al. 2007].

Each set of multiple features contains data extracted from different body parts. An example skeleton with 20 joints and their corresponding part-based multiple features are illustrated in Figure 1. As different actions may be well characterized by certain features of certain body parts, we divide each individual feature set into different groups according to different body parts and determine how discriminative they are for multi-task action classification. We then propose the Multiple Feature Sparse Fusion (MFSF) method by introducing joint group sparse regularization to learn the group sparse weight matrices of the sharable and specific feature structures. The proposed MFSF method can obtain the sharable structures among the part-based multiple features as well as the specific structures of part-based individual feature sets, both with group sparsity corresponding to different body parts.

The contributions of our work are stated as follows. First, our proposed Multiple Feature Sparse Fusion (MFSF) is a novel approach for human action recognition from depth video. The joint group sparse regularization is used in the learning stage to select discriminative sharable and specific structures among part-based multiple features for multi-task action classification. Second, since our MFSF model employs two non-smooth regularizers, we propose an efficient algorithm to solve for the optimization parameters. The resulting parameters can select key part-based features according to different types of actions.

The rest of this paper is organized as follows. We provide a brief review of the existing literature in Section 2 and give a framework overview in Section 3. We present the proposed learning scheme in Section 4. Experimental results are reported and analyzed in Section 5. We conclude the paper in Section 6.

*e-mail: mli269-c@my.cityu.edu.hk
†e-mail: howard@um.cityu.edu.hk
‡e-mail: hubert.shum@northumbria.ac.uk

MiG '16, October 10-12, 2016, Burlingame, CA, USA
ISBN: 978-1-4503-4592-7/16/10
DOI: http://dx.doi.org/10.1145/2994258.2994268

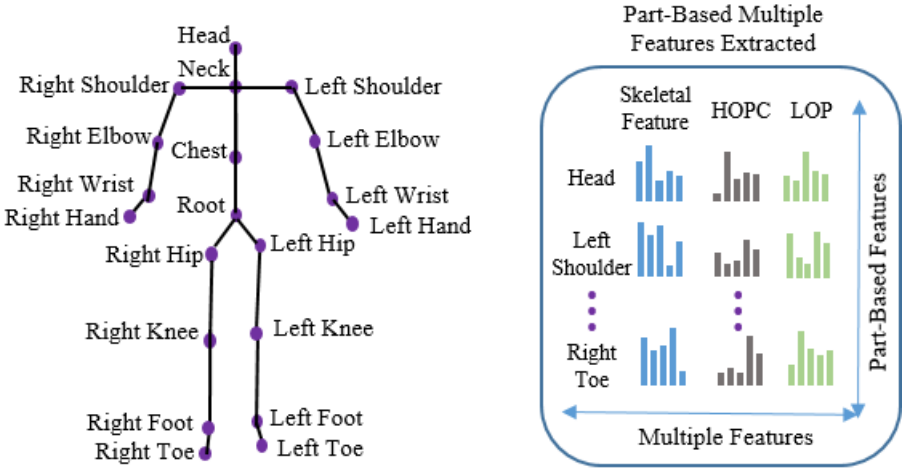

Figure 1: *Skeletal joint locations and their corresponding part-based multiple features.*

2 Related Work

Human actions captured in depth videos are high dimensional signals with huge spatio-temporal variations. The two major types of visual features extracted from depth signals are those inferred from the skeletal joints and those extracted directly from the depth map data.

The skeletal features are based on the 3D locations of skeleton joints on each frame of the action sequence, and they are very discriminative for action recognition. In [Yang and Tian 2014], "eigenjoints" were extracted from the 3D locations of joints for human action classification using the Naive-Bayes nearest neighbor rule. [Xia et al. 2012] utilized the spherical histograms of 3D locations of the joints with Hiddenv Markov Models (HMMs) to model the temporal changes of the skeletal action for classification task. However, the noise in estimated 3D locations of joints and occlusion of human body parts bound the reliability of this type of features. In addition, the 3D skeleton data alone is not sufficient to distinguish the interaction between the human subject and the surrounding objects.

Other than skeletal joint features, some features directly extract holistic or local descriptions from input depth map sequence. [Li et al. 2010] sampled boundary pixels from 2D silhouettes as a bag of features. [Yang et al. 2012] added the temporal derivative of 2D projections to get Depth Motion Maps (DMM). Space-Time Interest Point (STIP) detection described by Histogram of Oriented Gradients (HOG) [Dalal and Triggs 2005] and Histogram of Optical Flow (HOF) [Laptev 2005] were originally proposed for recognition tasks on RGB videos, but [Ni et al. 2013] showed that they could easily be generalized to handle RGB-D signals. Recently, [Oreifej and Liu 2013] extended histogram of oriented 3D normals [Klaser et al. 2008] to 4D (HON4D) by adding time derivative that was shown to be informative for action recognition. However, as is shown in [Rahmani et al. 2014a], information from very strong derivative locations, such as edges and silhouettes, may get suppressed [Rahmani et al. 2014b]. In order to improve the discrimination of descriptors, [Rahmani et al. 2014a] proposed the Histogram of Oriented Principal Components (HOPC) to capture the local geometric characteristics around each point within a sequence of 3D point clouds. The HOPC descriptor is more informative than HON4D as it captures the spread of data in three principal directions.

Since different features may have their own strength under various occasions, it is suggested to integrate them to encode human actions for recognition such that these multiple features can improve the discriminative power of human action representation. [Yu et al. 2014] integrated three types of features to construct a spatio-temporal representation, including pairwise joint distances, spatial joint coordinates, and temporal variations of joint locations. [Chaaraoui et al. 2013] applied feature fusion with skeletal and silhouette based features in order to obtain a visual feature for human action recognition. [Wang et al. 2014] defined "actionlet" as the combination of a limited numbers of joint features for action recognition. The aforementioned multiple feature fusion methods either just use skeletal features or simply concatenate different types of feature for multiple feature fusion. [Gao et al. 2015] applied multi-feature mapping and dictionary learning model to design human action recognition algorithms. [Liu et al. 2016] proposed a framework to fuse the depth map feature learned by a CNN model and skeletal feature for action recognition with depth sequences.

In this work, we extract multiple features consisted of skeletal joint features and two local depth appearance features (LOP and HOPC) from the depth videos. Moreover, we divide each individual feature set into different feature groups according to different body parts, and introduce group spare learning for the weights, which was not considered in the previous work. In particular, our proposed method learns the weights for the sharable and specific feature structures learnt among the part-based multiple features via the joint group sparse regularization. The resulting group sparse weight matrices help to select the discriminative part-based feature structures to improve multi-task action classification.

3 Framework Overview

In this section, we give a brief overview of our framework for human action recognition.

We first extract the skeletal feature, and two depth based features LOP and HOPC. Second, using these three types of part-based features, we propose the Multiple Feature Sparse Fusion (MFSF) to obtain the sharable and specific structures of the multiple features. Using the group sparse regularization, the proposed MFSF can rank the importance of the sharable and specific feature structures according to each body part. Third, we apply the weighted

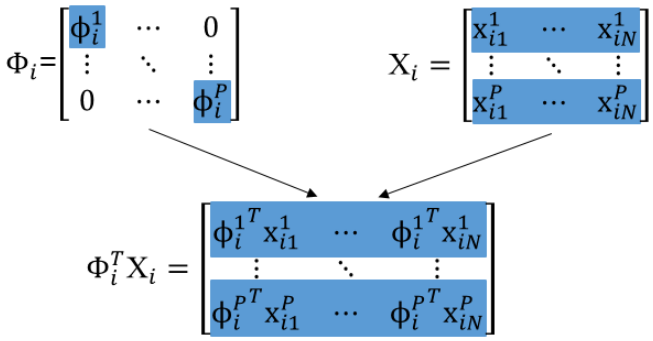

Figure 3: *Illustration of the calculation of* $\Phi_i{}^T X_i$.

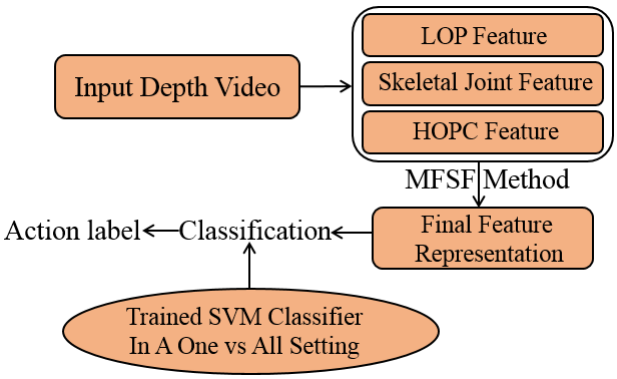

Figure 2: *Flow chart of our proposed approach.*

sharable and specific feature structures to construct the final representation for each action. In the end, we train the SVM classifier for action recognition using the final feature vectors. Figure 2 shows the flow chart of our proposed approach.

4 Multiple Feature Sparse Fusion

Recently, sparse regularization has been studied in a variety of research fields [Evgeniou and Pontil 2007; Wang et al. 2011; Wang et al. 2013]. Since group sparsity-inducing matrix norms can capture the group-wise importance of the elements in the matrices, in our learning framework, we will use joint group sparsity regularizers to learn the group sparse weight matrices for sharable and specific structures among part-based multiple features, in favor of selecting discriminative feature structures of certain body parts. In the following, we define our notation first, and then present a detailed description of the proposed MFSF.

4.1 Multiple feature extraction

The use of skeletal features is inspired by the work [Wang et al. 2014]. Following their practice, we repeatedly partition the temporal skeletal features into 1, 2, 4 sub-segments along the temporal dimension, and then concatenate the low frequency Fourier coefficients extracted from each segment. In addition to skeletal features, other types of features we use are local HOPC [Rahmani et al. 2014a] and LOP [Wang et al. 2014] to represent depth based local dynamics and appearance around each joint. LOP feature computes the local occupancy information based on the 3D point cloud around a particular skeletal joint, so that the temporal dynamics of all such local occupancy patterns can roughly discriminate different types of interactions. LOPs are extracted in a local region around each joint on each frame. For each joint on each frame, the local region is divided into $3 \times 3 \times 4$ numbers of bins, and the size of each bin is $6 \times 6 \times 80$ pixels. The LOP feature is computed in each bin, and then all the LOP features in all the bins are concatenated. Then, we use a similar Fourier temporal pyramid transformation to represent LOP features. HOPC features are also extracted locally over the location of joints on each frame. For each human skeletal joint on each frame, the local region is divided into $3 \times 3 \times 1$ numbers of bins, and the size of each bin is $12 \times 12 \times 6$ pixels. We then compute the HOPC histograms in all the bins. The concatenation of the HOPC histograms in all the bins are used as the final local descriptor.

4.2 The feature fusion model using joint group sparse regularization

In this subsection, we proposed the MFSF to obtain the sharable and specific structures of multiple features. The importance of sharable feature structure and specific features structure according to each body part is obtained by two group sparse regularization terms in the proposed MFSF.

Suppose there are M types of features. For each feature type i, let $\mathrm{X}_i = [\mathrm{x}_{i1}, ..., \mathrm{x}_{iN}] \in \Re^{d_i \times N}$ denotes the i-th type of feature matrix for N training samples, where d_i is the dimension of i-th type of feature. The feature set inside X_i is divided into P feature groups according to different body parts, and each $\mathrm{x}_{in} = [\mathrm{x}_{in}^1, ..., \mathrm{x}_{in}^P]^T \in \Re^{d_i}$ $(n \in [1, ..., N])$. We attempt to learn a projection matrix $\Phi_i \in \Re^{d_i \times S}$ (usually, $S \ll d_i$) with the j-th diagonal block as $\phi_i{}^j$ $(j \in [1, ..., P])$ for each X_i to project the P feature groups inside X_i into P subspaces spanned by the columns of the corresponding $\phi_i{}^j$ as illustrated in Figure 3. We have $M \times P$ subspaces, which are set to have the same dimensionality such that both the sharable and specific feature structures among part-based multiple features can be easily quantified in the subspaces by the weight matrices $\mathrm{W}_0 = [\mathrm{w}_{01}^1, ..., \mathrm{w}_{0C}^1; ..., ..., ...; \mathrm{w}_{01}^P, ..., \mathrm{w}_{0C}^P]$, $\mathrm{W}_i = [\mathrm{w}_{i1}^1, ..., \mathrm{w}_{iC}^1; ..., ..., ...; \mathrm{w}_{i1}^P, ..., \mathrm{w}_{iC}^P] \in \Re^{S \times C}$, where C indicates the number of action classes.

As illustrated in Figure 4, each w_{0c}^j $(c \in [1, ..., C])$ in W_i indicates the weight for sharable structures among all types of features corresponding to the j-th body part with respective to the c-th class, and each w_{ic}^j in W_i indicates the weight for specific structures of i-type of feature corresponding to the j-th body part with respective to the c-th class. We use $\mathrm{Y} \in \{-1, C-1\}^{C \times N}$ to represent the labels of all training samples, and define each column of Y as a zero-mean vector $[-1, ..., -1, C-1, -1, ..., -1]^T$. For a sample with class label c, the c^{th} entry of the zero-mean vector equals to a constant positive number $C - 1$.

To obtain the projects W_0^*, $\mathrm{W}_i^* \in \Re^{S \times C}$, $\Phi_i^* \in \Re^{d_i \times S}$ for mining

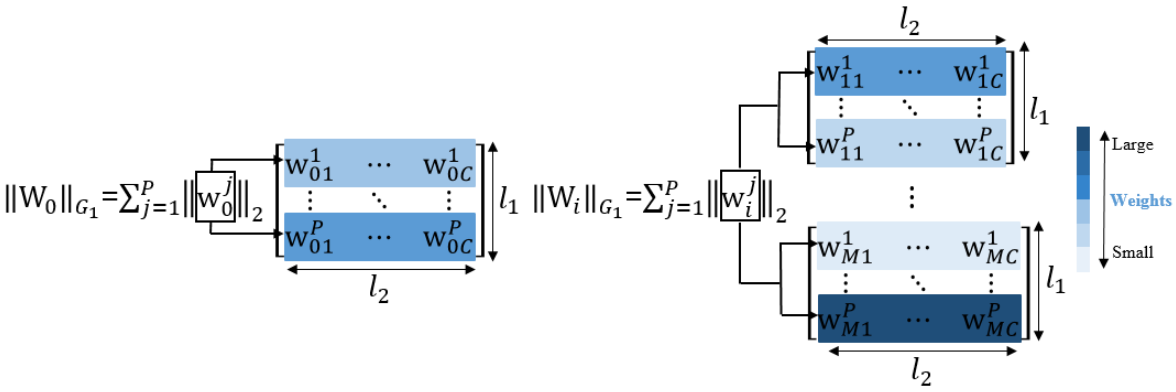

Figure 4: *Illustration of the weight matrices (i.e., W_0 and W_i) for sharable and specific feature structures. The rows in each weight matrix cluster into P groups according to the body parts. The weights in the row groups with deep blue color have large values. For example, we can see that the weights of M-th type feature of P-th body part have large values in this figure. The G_1-norm imposes group sparsity between different row groups in the weight matrices for sharable and specific feature structures.*

sharable and specific structure of multiple features, we propose the Multiple Feature Sparse Fusion (MFSF) model in the multi-task learning framework formulated as

$$\min_{\mathrm{W}_0,\{\mathrm{W}_i\},\{\Phi_i\}} \sum_{i=1}^{M} (J((\mathrm{W}_0+\mathrm{W}_i)^T\Phi_i^T\mathrm{X}_i,\mathrm{Y})+\beta\|\mathrm{W}_i\|_{G_1}+ \gamma J(\mathrm{X}_i,\Phi_i\Phi_i^T\mathrm{X}_i))+\alpha\|\mathrm{W}_0\|_{G_1} \quad (1)$$

$$\text{s.t. } \Phi_i^T\Phi_i=\mathrm{I}, i=1,2,...M.$$

The proposed MFSF can be represented as a least-square problem, where $J(A,B)=(A-B)^2$. The first term aims to jointly learn the common subspaces, and the sharable and specific feature structures. The third term intends to deliver good reconstruction of each X_i using projection matrix Φ_i. The $\Phi_i^T\Phi_i=\mathrm{I}$ is applied for reduction of redundant information. The G_1-norm in regularization terms $\|\mathrm{W}_0\|_{G_1}$ and $\|\mathrm{W}_i\|_{G_1}$ is group ℓ_1-norm [Yuan and Lin 2006]. $\|\mathrm{W}_0\|_{G_1}$ and $\|\mathrm{W}_i\|_{G_1}$ are defined as $\sum_{j=1}^{P}\|\mathrm{w}_0^j\|_2$ and $\sum_{j=1}^{P}\|\mathrm{w}_i^j\|_2$ (illustrated in Figure 3), where $\mathrm{w}_0^j=[\mathrm{w}_{01}^j,...,\mathrm{w}_{0C}^j]$, $\mathrm{w}_i^j=[\mathrm{w}_{i1}^j,...,\mathrm{w}_{iC}^j]$. G_1-norm uses ℓ_2-norm within sharable and specific feature structures corresponding to each body part and ℓ_1-norm between these structures. Hence, it enforces the sparsity between different sharable and specific structures, i.e., if sharable or specific feature structures for certain body part are not discriminative for multi-task classification, the objective in Eq. (1) will assign zeros (in ideal case, usually they are very small values) to them; otherwise, their weights are large. This norm regularizer emphasises the importance of different sharable and specific feature structures according to body parts. Thus, MFSF results automatically perform the selection procedure of these part-based feature structures.

4.3 Optimization algorithm

We introduce an alternative optimization scheme in three steps as follows.

Step 1. Fixing the coefficients W_i and Φ_i, optimize W_0:

$$\min_{\mathrm{W}_0}\sum_{i=1}^{M} J((\mathrm{W}_0+\mathrm{W}_i)^T\Phi_i^T\mathrm{X}_i,\mathrm{Y})+\alpha\|\mathrm{W}_0\|_{G_1}. \quad (2)$$

Then we can get[1]

$$\mathrm{W}_0=(\sum_{i=1}^{M}\Phi_i^T\mathrm{X}_i\mathrm{X}_i^T\Phi_i+\alpha\mathrm{D}_0)^{-1}\sum_{i=1}^{M}(\Phi_i^T\mathrm{X}_i(\mathrm{Y}^T-\mathrm{X}_i^T\Phi_i\mathrm{W}_i)). \quad (3)$$

D_0 is a block diagonal matrix with the j-th diagonal block as $\frac{1}{2\|\mathrm{w}_0^j\|_2}\mathrm{I}_j$, I_j is an identity matrix, w_0^j is the j-th segment of W_0 and includes the weights of sharable feature structures corresponding to the j-th body part. Note that D_0 is dependent on W_0 and thus is also unknown variable. We propose an iterative algorithm to solve this problem, which is described in Algorithm 1.

Algorithm 1 Our method for optimizing problem (Eq. (2)).

Input: $P,S,\alpha,\mathrm{X}_i\in\Re^{d_i\times n},\mathrm{Y}\in\Re^{C\times n}$.
Output: W_0

1: Let t=1. Initialize $\mathrm{W}_0(t)\in\mathbb{R}^{S\times C}$. Each ${\phi_i}^j$ in Φ_i is set as the top $\frac{S}{P}$ principal components of $[\mathrm{x}_{i1}^j,...,\mathrm{x}_{iN}^j]$ in X_i.
2: **while** not converge **do**
3: Calculate the block diagonal matrix $\mathrm{D}_0(t)$, where the j-th diagonal block of $\mathrm{D}_0(t)$ is $\frac{1}{2\|\mathrm{w}_0^j(t)\|_2}\mathrm{I}_j$.
4: For W_0, $\mathrm{W}_0(t+1)=(\sum_{i=1}^{M}\Phi_i^T\mathrm{X}_i\mathrm{X}_i^T\Phi_i+\alpha\mathrm{D}_0(t))^{-1}\sum_{i=1}^{M}(\Phi_i^T\mathrm{X}_i(\mathrm{Y}^T-\mathrm{X}_i^T\Phi_i\mathrm{W}_i))$.
5: $t=t+1$.
6: **end while**

Step 2. Fixing the coefficients W_0 and Φ_i, optimize W_i:

$$\min_{\{\mathrm{W}_i\}}\sum_{i=1}^{M}(J((\mathrm{W}_0+\mathrm{W}_i)^T\Phi_i^T\mathrm{X}_i,\mathrm{Y})+\beta\|\mathrm{W}_i\|_{G_1}). \quad (4)$$

Decoupling the problem above into the following independent group sparse-regularized unconstrained least square problems:

$$\min_{\mathrm{W}_i} J((\mathrm{W}_0+\mathrm{W}_i)^T\Phi_i^T\mathrm{X}_i,\mathrm{Y})+\beta\|\mathrm{W}_i\|_{G_1}. \quad (5)$$

[1]When $\|\mathrm{w}_0^j\|_2$=0, Eq. (2) is not differentiable. Following [Gorodnitsky and Rao 1997], we can introduce a small perturbation to regularize the j-th diagonal block of D_0 as $\frac{1}{2\sqrt{\|\mathrm{w}_0^j\|_2^2+\eta}}\mathrm{I}_j$. Then it can be verified that the derived algorithm minimizes the following problem: $\sum_{i=1}^{M}\|(\mathrm{W}_0+\mathrm{W}_i)\Phi_i^T\mathrm{X}_i-\mathrm{Y}\|_F^2+\alpha\sum_{j=1}^{P}\sqrt{\|\mathrm{w}_0{}^j\|_2^2+\eta}$, which is apparently reduced to problem Eq. (2) when $\eta\to 0$.

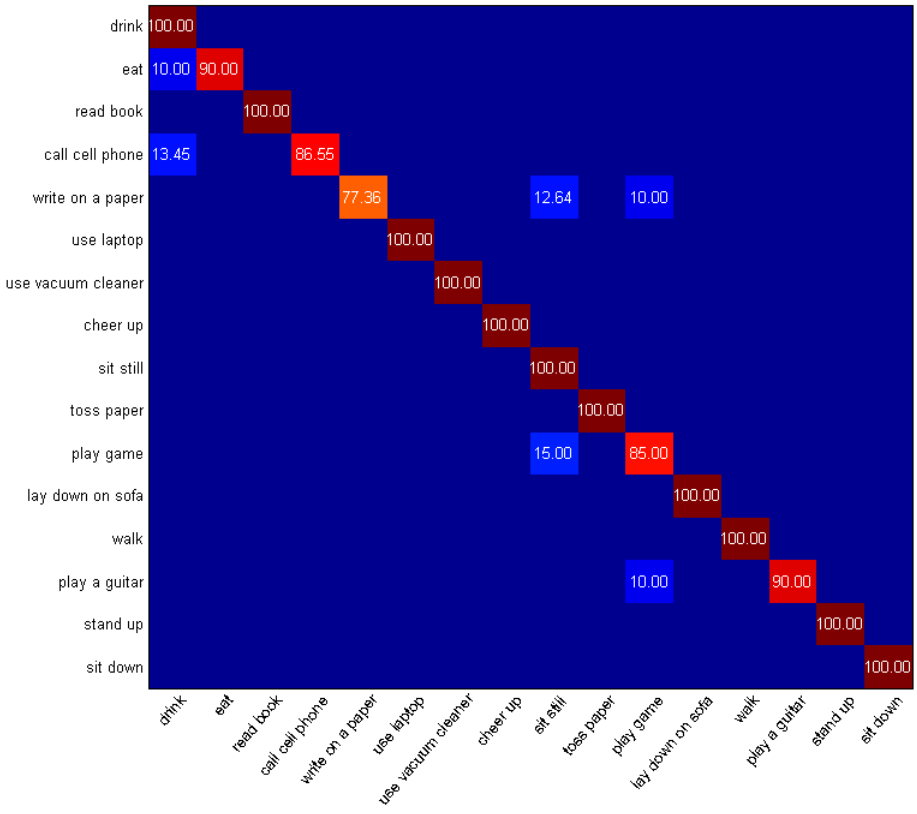

Figure 5: *Confusion matrix for MSR-Daily Activity dataset.*

Algorithm 2 Our method for optimizing problem (Eq. (4)).

Input: $P, S, \beta, \mathrm{X}_i \in \Re^{d_i \times n}, \mathrm{Y} \in \Re^{C \times n}$.
Output: W_i

1: Let t=1. Initialize $\mathrm{W}_i(t) \in \mathbb{R}^{S \times C}$. Each ${\phi_i}^j$ in Φ_i is set as the top $\frac{S}{P}$ principal components of $[\mathrm{x}_{i1}^j, ..., \mathrm{x}_{iN}^j]$ in X_i.
2: **while** not converge **do**
3: Calculate the block diagonal matrix $\mathrm{D}_i(t)$, where the j-th diagonal block of $\mathrm{D}_i(t)$ is $\frac{1}{2\|\mathrm{w}_i^j(t)\|_2}\mathrm{I}_j$.
4: For W_i, $\mathrm{W}_i(t+1) = (\Phi_i^T \mathrm{X}_i \mathrm{X}_i^T \Phi_i + \beta \mathrm{D}_i(t))^{-1} \Phi_i^T \mathrm{X}_i (\mathrm{Y}^T - \mathrm{X}_i^T \Phi_i \mathrm{W}_0)$.
5: $t = t + 1$
6: **end while**

Then we can obtain[2]

$$\mathrm{W}_i = (\Phi_i^T \mathrm{X}_i \mathrm{X}_i^T \Phi_i + \beta \mathrm{D}_i)^{-1} \Phi_i^T \mathrm{X}_i (\mathrm{Y}^T - \mathrm{X}_i^T \Phi_i \mathrm{W}_0). \quad (6)$$

D_i is a block diagonal matrix with the j-th diagonal block as $\frac{1}{2\|\mathrm{w}_i^j\|_2}\mathrm{I}_j$, w_i^j is the j-th segment of W_i and includes the weights of i-th specific feature structures corresponding to the j-th body part. Note that D_i is dependent on W_i and thus is also unknown variable. Hence, we propose an iterative algorithm which is described in Algorithm 2.

[2]When $\|\mathrm{w}_i^j\|_2$=0, Eq. (4) is not differentiable. Similarly as in footnote 1, we can regularize the j-th diagonal block of D_i as $\frac{1}{2\sqrt{\|\mathrm{w}_i^j\|_2^2+\eta}}\mathrm{I}_j$.

Step 3. Finally, we fix W_0, W_i, optimize Φ_i:

$$\min_{\{\Phi_i\}} \sum_{i=1}^{M} (J((\mathrm{W}_0 + \mathrm{W}_i)^T \Phi_i^T \mathrm{X}_i, \mathrm{Y}) + \gamma J(\mathrm{X}_i, \Phi_i \Phi_i^T \mathrm{X}_i)) \quad s.t. \Phi_i^T \Phi_i = \mathrm{I}, i = 1, 2, ...M. \quad (7)$$

In step 3, we follow [Wen and Yin 2013] to solve the Eq. (7). Given the $\Phi_i(t)$ in t-step, we first define a skew-symmetric matrix $\Theta = \nabla \Phi_i(t)^T - \Phi_i(t) \nabla^T$, in which ∇ is the gradient of the objective function, and can be indicated by $\nabla = \mathrm{X}_i((\mathrm{W}_0 + \mathrm{W}_i)^T \Phi_i(t)^T \mathrm{X}_i - \mathrm{Y})^T (\mathrm{W}_i + \mathrm{W}_0)^T - 2\gamma \mathrm{X}_i \mathrm{X}_i^T \Phi_i(t)$. Then the new updated point can be determined by the Grank-Nicolson-like scheme $\Phi_i(t+1) = (\mathrm{I} + \frac{\sigma}{2}\Theta)^{-1}(\mathrm{I} - \frac{\sigma}{2}\Theta)\Phi_i(t)$, in which σ is the iteration step size. In each iteration, optimal size would be determined by a line search method. We summarize our optimization for Eq. (1) in Algorithm 3.

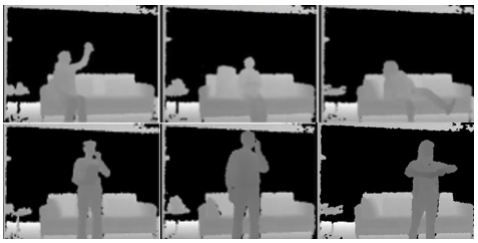

Figure 7: *Sample frames of the MSR-Daily Activity dataset.*

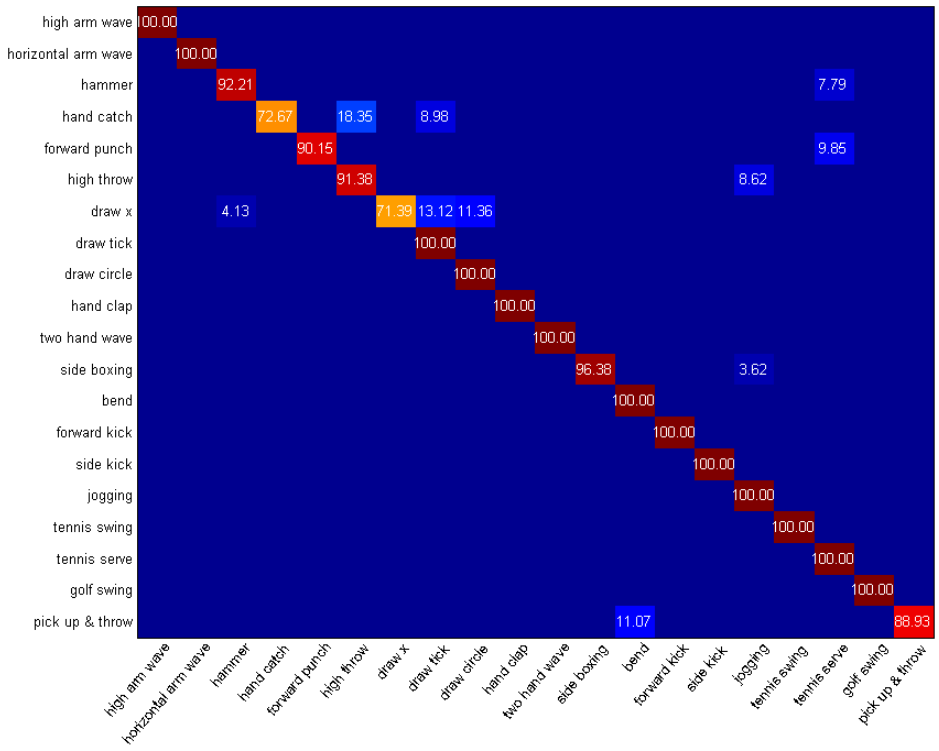

Figure 6: *Confusion matrix for MSR-Action 3D dataset.*

Algorithm 3 Our method for optimizing problem (Eq. (1)). Note that *Iter* and *maxIter* indicate the number variation of iteration and maximum number of iteration in Eq. (7).

Input: $P, S, \alpha, \beta, \gamma, maxIter, \mathrm{X}_i \in \Re^{d_i \times n}, \mathrm{Y} \in \Re^{C \times n}$.
Output: $\mathrm{W}_0, \mathrm{W}_i, \Phi_i$

1: Initialize $\mathrm{W}_0, \mathrm{W}_i \in \mathbb{R}^{S \times C}$. Each ${\phi_i}^j$ in Φ_i is set as the top $\frac{S}{P}$ principal components of $[\mathrm{x}_{i1}^j, ..., \mathrm{x}_{iN}^j]$ in X_i. *Iter* $= 1$.
2: **while** *Iter* $<$ *maxIter* **do**
3: $\quad \mathrm{W}_0 \leftarrow$ Output of Algorithm 1.
4: $\quad \mathrm{W}_i \leftarrow$ Output of Algorithm 2, $i = 1, 2, ...M$
5: $\quad$ **for** $i = 1$ *to* M **do**
6: $\quad\quad \nabla \leftarrow \mathrm{X}_i((\mathrm{W}_0 + \mathrm{W}_i)^T \Phi_i^T \mathrm{X}_i - \mathrm{Y})^T (\mathrm{W}_i + \mathrm{W}_0)^T - 2\gamma \mathrm{X}_i \mathrm{X}_i^T \Phi_i$.
7: $\quad\quad \Theta \leftarrow \nabla \Phi_i^T - \Phi_i \nabla^T$.
8: $\quad\quad \Phi_i \leftarrow (\mathrm{I} + \frac{\sigma}{2}\Theta)^{-1}(\mathrm{I} - \frac{\sigma}{2}\Theta)\Phi_i$.
9: $\quad$ **end for**
10: $\quad$ *Iter*=*Iter*+1
11: **end while**

4.4 Construction of final feature representation

Using the learnt parameters $\mathrm{W}_0, \mathrm{W}_i$ and Φ_i, we first define two confidence vectors to encode the shared and specific feature structures of each new sample $\mathrm{x}_i, i = 1, 2...M$ with part-based multiple features $\mathrm{V}_{sharable}^i = \mathrm{W}_0^T \Phi_i^T \mathrm{x}_i \in \Re^C$ and $\mathrm{V}_{specific}^i = \mathrm{W}_i^T \Phi_i^T \mathrm{x}_i \in \Re^C$. Inspired by the augmented feature construction in [Li et al. 2014], we concatenate all the sharable confidence vectors and all the specific confidence vectors together to form higher-level augmented features and combine these augmented features to construct our final representation. Then, using the final feature, we train a linear SVM classifier to make the final classification decision.

5 Experiments

In this section, we evaluated our methods on three human action recognition benchmarks. In all our experiments, we use LIBSVM software [Chang and Lin 2011] with our final feature description to train our linear SVM classifier. In the following, we first briefly introduce the implementation details and then describe the experiments and results.

5.1 Implementation datails

All the experiments are done on Kinect-based datasets. The outputs of Kinect are multiple signals that give RGB videos, depth sequences and skeletal information. In order to have a fair comparison with other depth based methods, we ignore the RGB signal. For the MFSF model, there are three parameters: α, β and γ that corresponding to group sparsity (α, β) and reconstruction loss (γ), respectively. According to our observation, the performance is best when $\alpha = 0.1 \sim 0.2$, $\beta = 0.1 \sim 0.2$ and $\gamma = 1 \sim 2$. For

the dimensionality S of the subspace, we set its value about $\frac{1}{3}$ of the number of the training samples to obtain the stable accuracy results. 30 iterations ($maxIter$ in Algorithm 3) are set for obtaining a reliable solution in all of our experiments.

5.2 MSR-Daily Activity dataset

According to its intra-class variations and choices of action classes, MSR-Daily Activity dataset [Wang et al. 2014] is one of the most challenging benchmarks for human action recognition. This dataset contains 16 types of activities: *drink*, *eat*, *read book*, *call cell phone*, *write on a paper*, *use laptop*, *use vacuum cleaner*, *cheer up*, *sit still*, *toss paper*, *play game*, *lie down on sofa*, *walk*, *play a guitar*, *stand up*, *sit down*. A skeleton has 20 joint positions. The total number of the activity samples is 320. Most of the activities involve human-objective interactions. We follow the same experiment setting as other related works, where half of the subjects are used as training data, while the other half are used as testing data.

Table 1: *Comparative results on MSR-Daily Activity dataset using different norms.*

Method	Different norms used	Accuracy
F squared, F squared	$\|\mathrm{W}_0\|_F^2$, $\|\mathrm{W}_i\|_F^2$	92.5
G_1, F squared,	$\|\mathrm{W}_0\|_{G_1}$, $\|\mathrm{W}_i\|_F^2$	93.2
F squared, G_1	$\|\mathrm{W}_0\|_F^2$, $\|\mathrm{W}_i\|_{G_1}$	93.9
MFSF	$\|\mathrm{W}_0\|_{G_1}$, $\|\mathrm{W}_i\|_{G_1}$	**95.6**

We first assess the power of our joint group sparse regularization by evaluating the performance of different joint regularization using the plain Frobenius norm and the group ℓ_1-norm for part-based multiple feature fusion in the multi-task action recognition. The results of this experiment are shown in Table 1. It can be seen that, using group ℓ_1-norm can improve the accuracy performance. Especially in the case of MFSF, the improvement is more significant, and accuracy rise by more than 3% comparing with the case of just using plain Frobenius norm squared.

Table 2: *Comparative results on MSR-Daily Activity dataset based on single type of features.*

Method	Types	Accuracy
Only LOP feature [Wang et al. 2014]	LOP	42.5
Proposed MFSF	LOP	**67.3**
Actionlet [Wang et al. 2014]	Skeleton	68
Proposed MFSF	Skeleton	**80.2**
Local HOPC [Rahmani et al. 2014a]	HOPC	81.7
Proposed MFSF	HOPC	**83.5**

Then, we verify our method in the case of single-type features without mining sharable structures among part-based multiple features. As shown in Table 2, using LOP features, we achieve 67.3% compared to 42.5% of the actionlet method. For skeletal based features, we obtain 80.2% which is more than 12% higher than the performance of actionlet. On local HOPC features, we reach 83.5% compared to 81.7% of the local HOPC method.

To verify the strength of the proposed MFSF, we try the different combinations of multiple features. As provided in Table 3, using skeletal and LOP features, we get 90.0% of accuracy which outperforms the 85.8% of actionlet method. And finally, using all three types, MFSF achieve the best performance of 95.6%. Comparing with the results of SFSL in Table 2, it also shows the benefit of the selected multiple feature fusion.

Table 3: *Comparative results on MSR-Daily Activity dataset based on combinations of multiple features.*

Method	Types	Accuracy
Actionlet [Wang et al. 2014]	Skeleton+LOP	85.8
Proposed MFSF	Skeleton+LOP	**90.0**
Proposed MFSF	HOPC+LOP	**91.3**
Proposed MFSF	Skeleton+HOPC	**92.8**
Proposed MFSF	Skeleton+LOP+HOPC	**95.6**

The confusion matrix of the results by our MFSF method is presented in Figure 5. It is clear that our method achieves perfect classification results on 11 action classes. The larger error is due to the misclassifications of the actions of 'call cell phone' as 'drink' and 'play (electronic) game' as 'sit still'. The reason lies in the high similarities between each pair of actions.

5.3 MSR-Action 3D dataset

Figure 9: *Sample frames of the MSR-Action 3D dataset.*

MSR-Action 3D dataset [Li et al. 2010] was captured using a depth sensor similar to Kinect. A skeleton has 20 joint positions. It consists of 20 actions: *high arm wave, horizontal arm wave, hammer, hand catch, forward punch, high throw, draw x, draw tick, draw circle, hand clap, two hand wave, side boxing, bend, forward kick, side kick, jogging, tennis swing, tennis serve, golf swing, pick up & throw.*

Table 4: *Comparative results on MSR-Action 3D dataset.*

Method	Accuracy
Depth HOG [Yang et al. 2012] (as reported in [Xia and Aggarwal 2013])	85.5
Actionlet [Wang et al. 2014]	88.2
HON4D [Oreifej and Liu 2013]	88.9
DSTIP [Xia and Aggarwal 2013]	89.3
Lie Group [Vemulapalli et al. 2014]	89.5
3D^{2}CNN [Liu et al. 2016]	90.18
HOPC [Rahmani et al. 2014a]	91.6
Max Margin Time Warping [Wang and Wu 2013]	92.7
MMDLM [Gao et al. 2015]	93
Proposed MFSF	**94.3**

Each action was performed by ten subjects for three times. The frame rate is 15 frames per second and resolution 640×480. Altogether, the dataset has 23797 frames of depth map for 402 action

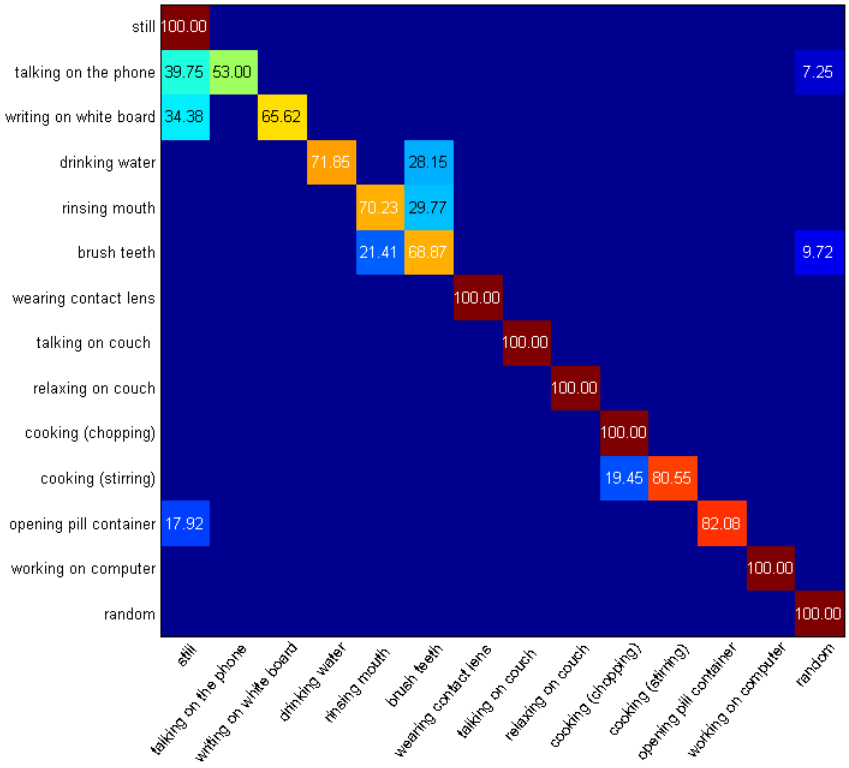

Figure 8: *Confusion matrix for CAD 60 dataset.*

performed by 10 different subjects. Each subject performed every action two or three times. Altogether, there are 557 action sequences. This is a challenging dataset because many of the actions are highly similar to each other.

We follow the same experiment setting as other state-of-the-art methods, where half of the subjects are used as test data, and the rest of the samples are used as test data, it is what other state-of-the-art methods also done in the comparison. As shown in Table 4, we still achieve the highest recognition accuracy among the reported results, however, the achieved margin is not as large as other datasets. This is because that there is not any interaction with other objectives, most of the classes are highly distinguishable just using the skeletal features. Therefore, our multiple feature combination could not boost up the accuracy results that much, but the group sparse learning still shows its advantage over other methods.

The confusion matrix is illustrated in Figure 6. It is clear that, our method works very well for most of the actions. The misclassifications occur if two actions are too similar to distinguish just using the skeletal features, such as 'hand catch' and 'high throw', or if the occlusion is so large that the 3D positions of the tracked joints are frequently inaccurate, such as the action 'pick up and throw'.

5.4 Cornell Activity dataset 60 (CAD 60)

Cornell Activity dataset (CAD-60) [Sung et al. 2011] contains 68 video clips captured with Kinect cameras. A skeleton has 15 joint positions. The actions in this dataset can be categorized into 5 different environments: office, kitchen, bedroom, bathroom, and living room. Three or four common activities were identified for each environment, giving a total of 13 specific actions: *still, talking on the phone, writing on white board, drinking water, rinsing mouth , brush teeth, wearing contact lens, talking on couch , relaxing on couch, cooking (chopping), cooking (stirring), opening pill container, working on computer.*

Table 5: *Comparative results on CAD 60 dataset .*

Method	Accuracy
STIP [Zhu et al. 2014]	62.5
Order Sparse Coding [Ni et al. 2012]	65.3
Local HOPC [Rahmani et al. 2014a]	73.5
Actionlet [Wang et al. 2014]	74.7
Hierarchical HMM [Raman and Maybank 2016]	85.4
Proposed MFSF	**86.9**

In this dataset, we use a more challenging experimental setting for a more effective comparison among different action recognition methods. We follow the same experimental setting as in [Wang et al. 2014] by adopting the leave-one-person-out cross-validation per environment, which ensures that person participating in the training cannot be seen in the testing. As shown in Table 5, the proposed method achieves an accuracy of 86.9%, which is better than the reported results of the state-of-the-art methods.

Figure 10: *Sample frames of the CAD 60 dataset.*

The confusion matrix of the results by our method is presented in Figure 8. It is clear that our method can achieve good performance in recognizing most of the actions. The the misclassifications occur when distinguishing action still from those actions with subtle motions (e.g., 'talking on the phone', 'writing on white board') or if two actions are too similar (e.g., 'rinsing mouth' and 'brush teeth').

6 Conclusion and Future Work

This paper presents a novel method called Multiple Feature Sparse Fusion (MFSF) model to fuse the part-based multiple features for action classification in depth sequences. Our MFSF method learns the weight matrices for the part-based sharable feature structures and specific feature structures, respectively, via the group sparse regularization. The natural property of the proposed joint group sparse regularization automatically identifies the important part-based sharable and specific feature structures. State-of-the-art results are achieved on three challenging depth based action recognition datasets, which shows the effectiveness of the proposed method.

Future work includes the exploring the application of our MFSF method for person-person interaction recognition, where how to determine the importance of different sharable and specific structures among part-based multiple features involving the person-person interactions should be considered.

Acknowledgment

The work described in this paper was supported by City University of Hong Kong (Project No. 7004548) and the Engineering and Physical Sciences Research Council (EPSRC) (Ref: EP/M002632/1).

References

AGGARWAL, J. K., AND XIA, L. 2014. Human activity recognition from 3d data: A review. *Pattern Recognition Letters 48*, 70–80.

AMIT, Y., FINK, M., SREBRO, N., AND ULLMAN, S. 2007. Uncovering shared structures in multiclass classification. In *Proceedings of the 24th international conference on Machine learning*, ACM, 17–24.

CHAARAOUI, A., PADILLA-LOPEZ, J., AND FLÓREZ-REVUELTA, F. 2013. Fusion of skeletal and silhouette-based features for human action recognition with rgb-d devices. In *Proceedings of the IEEE international conference on computer vision workshops*, 91–97.

CHANG, C.-C., AND LIN, C.-J. 2011. Libsvm: a library for support vector machines. *ACM Transactions on Intelligent Systems and Technology (TIST) 2*, 3, 27.

CHEN, J., TANG, L., LIU, J., AND YE, J. 2013. A convex formulation for learning a shared predictive structure from multiple tasks. *IEEE transactions on pattern analysis and machine intelligence 35*, 5, 1025–1038.

DALAL, N., AND TRIGGS, B. 2005. Histograms of oriented gradients for human detection. In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, vol. 1, IEEE, 886–893.

DU, Y., WANG, W., AND WANG, L. 2015. Hierarchical recurrent neural network for skeleton based action recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 1110–1118.

EVGENIOU, A., AND PONTIL, M. 2007. Multi-task feature learning. *Advances in neural information processing systems 19*, 41.

GAO, Z., ZHANG, H., LIU, A. A., XU, G., AND XUE, Y. 2015. Human action recognition on depth dataset. *Neural Computing and Applications*, 1–8.

GORODNITSKY, I. F., AND RAO, B. D. 1997. Sparse signal reconstruction from limited data using focuss: A re-weighted minimum norm algorithm. *IEEE Transactions on signal processing 45*, 3, 600–616.

HAN, F., REILY, B., HOFF, W., AND ZHANG, H. 2016. space-time representation of people based on 3d skeletal data: a review. *arXiv preprint arXiv:1601.01006*.

HO, E. S., CHAN, J. C., CHAN, D. C., SHUM, H. P., CHEUNG, Y.-M., AND YUEN, P. C. 2016. Improving posture classification accuracy for depth sensor-based human activity monitoring in smart environments. *Computer Vision and Image Understanding 148*, 97–110.

KLASER, A., MARSZAŁEK, M., AND SCHMID, C. 2008. A spatio-temporal descriptor based on 3d-gradients. In *BMVC 2008-19th British Machine Vision Conference*, British Machine Vision Association, 275–1.

LAPTEV, I. 2005. On space-time interest points. *International Journal of Computer Vision 64*, 2-3, 107–123.

LI, W., ZHANG, Z., AND LIU, Z. 2010. Action recognition based on a bag of 3d points. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition-Workshops*, IEEE, 9–14.

LI, W., DUAN, L., XU, D., AND TSANG, I. W. 2014. Learning with augmented features for supervised and semi-supervised heterogeneous domain adaptation. *IEEE transactions on pattern analysis and machine intelligence 36*, 6, 1134–1148.

LIU, Z., ZHANG, C., AND TIAN, Y. 2016. 3d-based deep convolutional neural network for action recognition with depth sequences. *Image and Vision Computing*.

NI, B., MOULIN, P., AND YAN, S. 2012. Order-preserving sparse coding for sequence classification. In *Computer Vision–ECCV 2012*. Springer, 173–187.

NI, B., WANG, G., AND MOULIN, P. 2013. Rgbd-hudaact: A color-depth video database for human daily activity recognition. In *Consumer Depth Cameras for Computer Vision*. Springer, 193–208.

OREIFEJ, O., AND LIU, Z. 2013. Hon4d: Histogram of oriented 4d normals for activity recognition from depth sequences. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 716–723.

PADILLA-LÓPEZ, J. R., CHAARAOUI, A. A., AND FLÓREZ-REVUELTA, F. 2014. A discussion on the validation tests employed to compare human action recognition methods using the msr action3d dataset. *arXiv preprint arXiv:1407.7390*.

RAHMANI, H., MAHMOOD, A., HUYNH, D. Q., AND MIAN, A. 2014. Hopc: Histogram of oriented principal components of 3d pointclouds for action recognition. In *European Conference on Computer Vision*, Springer, 742–757.

RAHMANI, H., MAHMOOD, A., HUYNH, D. Q., AND MIAN, A. 2014. Real time action recognition using histograms of depth gradients and random decision forests. In *IEEE Winter*

Conference on Applications of Computer Vision, IEEE, 626–633.

RAMAN, N., AND MAYBANK, S. 2016. Activity recognition using a supervised non-parametric hierarchical hmm. *Neurocomputing 199*, 163–177.

SHOTTON, J., SHARP, T., KIPMAN, A., FITZGIBBON, A., FINOCCHIO, M., BLAKE, A., COOK, M., AND MOORE, R. 2013. Real-time human pose recognition in parts from single depth images. *Communications of the ACM 56*, 1, 116–124.

SUNG, J., PONCE, C., SELMAN, B., AND SAXENA, A. 2011. Human activity detection from rgbd images. *plan, activity, and intent recognition 64*.

TAHA, A., ZAYED, H. H., KHALIFA, M., AND EL-HORBATY, E.-S. M. 2015. Skeleton-based human activity recognition for video surveillance. *International Journal of Scientific & Engineering Research 6*, 1.

TORRALBA, A., MURPHY, K. P., AND FREEMAN, W. T. 2007. Sharing visual features for multiclass and multiview object detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence 29*, 5, 854–869.

VEMULAPALLI, R., ARRATE, F., AND CHELLAPPA, R. 2014. Human action recognition by representing 3d skeletons as points in a lie group. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 588–595.

WANG, J., AND WU, Y. 2013. Learning maximum margin temporal warping for action recognition. In *Proceedings of the IEEE International Conference on Computer Vision*, 2688–2695.

WANG, H., NIE, F., HUANG, H., RISACHER, S., DING, C., SAYKIN, A. J., SHEN, L., ET AL. 2011. Sparse multi-task regression and feature selection to identify brain imaging predictors for memory performance. In *2011 International Conference on Computer Vision*, IEEE, 557–562.

WANG, H., NIE, F., HUANG, H., AND DING, C. 2013. Heterogeneous visual features fusion via sparse multimodal machine. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 3097–3102.

WANG, J., LIU, Z., AND WU, Y. 2014. Learning actionlet ensemble for 3d human action recognition. In *Human Action Recognition with Depth Cameras*. Springer, 11–40.

WEN, Z., AND YIN, W. 2013. A feasible method for optimization with orthogonality constraints. *Mathematical Programming 142*, 1-2, 397–434.

XIA, L., AND AGGARWAL, J. 2013. Spatio-temporal depth cuboid similarity feature for activity recognition using depth camera. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2834–2841.

XIA, L., CHEN, C.-C., AND AGGARWAL, J. 2012. View invariant human action recognition using histograms of 3d joints. In *2012 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, IEEE, 20–27.

YANG, X., AND TIAN, Y. 2014. Effective 3d action recognition using eigenjoints. *Journal of Visual Communication and Image Representation 25*, 1, 2–11.

YANG, X., ZHANG, C., AND TIAN, Y. 2012. Recognizing actions using depth motion maps-based histograms of oriented gradients. In *Proceedings of the 20th ACM international conference on Multimedia*, ACM, 1057–1060.

YU, G., LIU, Z., AND YUAN, J. 2014. Discriminative orderlet mining for real-time recognition of human-object interaction. In *Asian Conference on Computer Vision*, Springer, 50–65.

YUAN, M., AND LIN, Y. 2006. Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 68*, 1, 49–67.

ZHANG, Y., AND YEUNG, D.-Y. 2012. A convex formulation for learning task relationships in multi-task learning. *arXiv preprint arXiv:1203.3536*.

ZHU, Y., CHEN, W., AND GUO, G. 2014. Evaluating spatiotemporal interest point features for depth-based action recognition. *Image and Vision Computing 32*, 8, 453–464.