Jointly Learning Heterogeneous Features for RGB-D Activity Recognition

Jian-Fang Hu[†], Wei-Shi Zheng[‡⋆*], Jianhuang Lai[‡], and Jianguo Zhang[◇]
[†]School of Mathematics and Computational Science, Sun Yat-sen University, China
[‡]School of Information Science and Technology, Sun Yat-sen University, China
[⋆]Guangdong Province Key Laboratory of Computational Science, Guangzhou, China
[◇]School of Computing, University of Dundee, United Kingdom
hujianf@mail2.sysu.edu.cn, wszheng@ieee.org, stsljh@mail.sysu.edu.cn, jgzhang@computing.dundee.ac.uk

Abstract

In this paper, we focus on heterogeneous feature learning for RGB-D activity recognition. Considering that features from different channels could share some similar hidden structures, we propose a joint learning model to simultaneously explore the shared and feature-specific components as an instance of heterogenous multi-task learning. The proposed model in an unified framework is capable of: 1) jointly mining a set of subspaces with the same dimensionality to enable the multi-task classifier learning, and 2) meanwhile, quantifying the shared and feature-specific components of features in the subspaces. To efficiently train the joint model, a three-step iterative optimization algorithm is proposed, followed by two inference models. Extensive results on three activity datasets have demonstrated the efficacy of the proposed method. In addition, a novel RGB-D activity dataset focusing on human-object interaction is collected for evaluating the proposed method, which will be made available to the community for RGB-D activity benchmarking and analysis.

1. Introduction

The emergence of low-cost depth sensors (e.g., the Microsoft Kinect) opens a new dimension to address the challenge of human activity recognition. Compared to the conventional use of RGB videos, the information from depth channel is insensitive to illumination variations, invariant to color and texture changes, and more importantly reliable for estimating body silhouette and skeleton (human posture) [24]. Bearing on these merits, there are two emerging branches of activity recognition work: 1) depth-based representation and 2) RGB-D based development.

On building depth-based representation, a straightforward way is to generalize the descriptors developed for RGB channel to depth channel, typically for describing the shape geometry [13, 28, 16, 41, 22, 36, 16], where Depth HOG [22, 36] is one representative. Due to the development of realtime human skeleton tracker from single depth image [24], motion information could be effectively encoded using the positional dynamics of each individual skeletal joints [9, 33, 18] or relationship of joint pairs [35, 17, 21] or even their combination [40, 14, 27]. Recently, local depth patterns around each key joint of human skeleton are also found to be useful in [29, 30, 10].

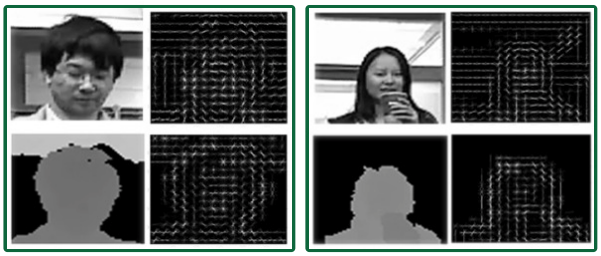


Figure 1. Visualization of HOG features from two activity snapshots in RGB (gray) and depth channel respectively. As shown, the HOG features from both channels of the same activity unveils similar "gist" structure of that activity, e.g., the"gist" of looking down in reading, and cup-to-mouth in drinking.

Depth does not necessarily mean discriminant. Albeit invariant to appearance changes, it does lose some useful information such as texture context, which is critical to distinguish some activities involving human-object interactions. Therefore, the RGB-D approach becomes an effective way to describe activities of interactions [30, 10, 15, 23, 4, 11]. For instance, Liu and Shao [15] simultaneously fused the RGB and depth information using a deep architecture; Zhu et al. [40] employed a set of random forests to fuse spatiotemporal and joints features; Shahroudy et al. [23] selected to fuse the RGB information and skeleton cues using

*corresponding author

a structured sparsity method.

However, the majority of these methods neither seek to jointly learn the features extracted from RGB and depth channels simultaneously nor model their underlying connections. As can be observed from Figure 1, features from different channels indeed share visual structures. The benefits of exploring both shared and specific structures for classification have been demonstrated by multi-task learning [5, 38, 2, 1], since it can significantly reduce the effective complexity of the task and transfer knowledge between related tasks [1, 26]. However, these methods assume that the features employed by different tasks are homogeneous, thus not applicable for mining shared and feature-specific structures among heterogeneous features.

In this paper, we propose a heterogeneous feature learning model for RGB-D activity recognition. Our model is built on mining a set of subspaces (one subspace for each heterogeneous feature) such that features with different dimensionality can be compared, and their shared and specific components can be easily encoded. We introduce one projection matrix as a linear transformation for each feature, and then formulate our subspaces mining and shared features learning in the framework of multi-task learning. Therefore, the optimal solutions for projection matrices and shared-specific structures can be jointly derived. Moreover, a three-step iterative optimization algorithm is proposed to find the optimal solution. We call the proposed model the **j**oint hete**r**ogeneous feat**u**res **le**arning (JOULE) model.

It is noted that there has been progress of multi-task learning for heterogeneous features with different assumptions in different context [7, 34, 39, 3]. Different features are combined in [3] by fusing of different metrics for recognition, but their shared latent structures are not considered. Realizing that different features may share some structures, the work of [34] assumes that different tasks share a *common set* of input variables (i.e., a common set of input features). However, this is not the case for our RGB-D based activity recognition, since our features are of different types with different dimensionality. The multi-task discriminant analysis (MTDA) is the closest to ours [39]. Our model is notably different from theirs, though both models unitize the concept of subspaces. Their model assumes there is a shared common space after projecting each type of features separately without explicitly considering the specific structures of each feature type. In contrast, we relax the assumption and assume heterogeneous features only partially share even after projection, which makes our method more applicable for describing the complex connections (shared and specific structures) among heterogeneous features extracted from RGB, depth and skeleton channels with large variations. In this context, we cast our model as a Frobenious-regularised least-square problem, with both *prediction* and *reconstruction* loss considered in an unified framework. This leads to a better overall performance of our model in the experiments.

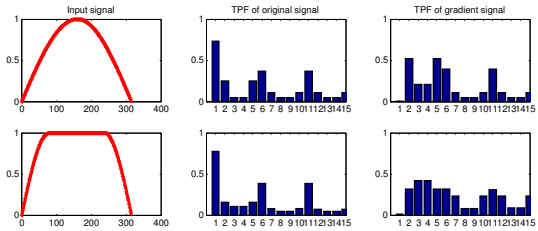


Figure 2. Two signals (left) and their TPF features (middle and right). The TPF features of the gradient signal (right) is more distinctive than the TPF of the original signal (middle) when differentiating the input signals (left).

In addition to a joint heterogeneous learning model, we propose to extend the temporal pyramid Fourier features (TPF) developed in [29] to both the original feature signal and its gradient to implicitly encode human motions, which experimentally yield better performance than TPF on original feature signal only.

To test the cross-subject generalization performance of our method on 3D human-object interactions more extensively, we also contribute a new RGB-D activity dataset containing 12 activity classes from 40 participants. Both this dataset and our code will be released to the public for benchmarking.

In summary, the contributions of our work are:1) a joint heterogeneous feature learning framework for RGB-D activity recognition; 2) an improved gradient temporal pyramid Fourier features; 3) a novel dataset collected for 3D human-object interaction recognition.

2. Heterogeneous Features Construction

We describe here in detail three descriptors utilized in our model: dynamic *skeleton* (DS) features, dynamic *color* pattern (DCP) and dynamic *depth* patterns (DDP). Each descriptor consists of two components: temporal pyramid Fourier features (TPF) from: i) the original feature signal and ii) the corresponding gradient signal respectively. These six components form our heterogeneous feature set.

The use of TPF features is motivated from the work of Wang et al. [29]. Following their practice, we repeatedly partition the feature signal (e.g., temporal skeleton features in [29]) into 1, 2 and 4 sub-segments along the temporal dimension, and then concatenate the low frequency Fourier coefficients extracted from each segment.
In addition to computing TPF from the original feature series as in [29], we also calculate TPF from the temporal gradient signal of the original feature series. This proposed extension is motivated from the following observations: 1) the gradient could, to a certain extent, implicitly encode the velocity change of the motion in activity; 2) it could also

capture the variation of pixel values, which helps describing the interactions between human and objects. For instance, the rapid change of the pixel values near a mouth may indicate some objects are coming near and interacting with the mouth (e.g., drinking). As illustrated in Figure 2, the temporal pyramid Fourier features of the gradient signal may capture more discriminative cues.

Dynamic Skeleton. Human pose and its dynamics are one of the key elements in activities [37, 8]. Here we extract the pose dynamics using skeleton information from the depth sequences for our activity modeling. Specifically, for each video sequence, the real-time skeleton tracker [24] is used to extract the trajectories of human key joints (skeleton). We then compute the relative positions of each trajectory pair as it was shown a discriminative feature for distinguishing different actions in [29]. The temporal pyramid Fourier features are further extracted from the relative positions as well as its gradient version to represent the dynamic pose information. It was noted that the sequence length may vary from video to video. Relative positions of each trajectory pair are interpolated by cubic spine to have the same length before computing the Fourier features, which ensures that the frequency locations of computed TPF features are properly calibrated and aligned before comparison.

Dynamic Color and Depth Pattern. Using the 3D joint positions without local appearance is often insufficient to characterize complex activities including human-object interactions. To compensate this, the local appearance features (both in RGB and depth) are extracted around each human joint, which could capture characteristic shape, texture and manipulated object's appearance during interactions. Specifically, for each joint in a trajectory, we first compute the HOG feature [6] in its local region (RGB or Depth) for all the associated frames. All of the HOG features of one joint trajectory constitute a temporal HOG tube. Then for each temporal dimension of the HOG tube, we extract the TPF features including the original and gradient version, and then concatenate them together to form our final descriptor. The HOG-TPF extracted from RGB sequence and depth sequence form our dynamic *color* pattern (DCP) and dynamic *depth* pattern (DDP), respectively.

3. Heterogeneous Feature Learning

Different features may share some similar structural components as illustrated in Figure 1. To effectively quantify the shared structures among different features with varied dimensions, we introduce a set of subspaces to represent these features so that they can be compared directly. These subspaces are learned by balancing the trade-off between the shared structures and feature-specific cues. In the following, we define our notations first, and then present a detailed description of the proposed joint learning model.

3.1. The Joint Learning Model

Suppose there are S types of heterogeneous features to learn together. For each feature type i $(i = 1, ..., S)$, let $\mathbf{X}_i \in \mathbb{R}^{d_i \times n}$ denote the feature matrix of n training instances, where d_i represents the feature dimensionality. We attempt to learn a projection matrix $\mathbf{\Theta}_i$ for each X_i to project it into a subspace spanned by the columns of $\mathbf{\Theta}_i$. In total, we have S subspaces, which are set to have the same dimensionality such that both the shared and feature-specific structures across different feature types can be easily quantified in the subspaces by two weight matrix $\mathbf{W}_0, \mathbf{W}_i \in \mathbb{R}^{M \times L}$, where M is the dimensionality of the subspace, and usually $M << d_i$. L indicates the number of activity classes. We use $\mathbf{Y} \in \{-1, L-1\}^{L \times n}$ to represent the labels of all the training samples. Each column of Y is defined as a zero-mean vector $[-1, ..., -1, L-1, -1..., -1]^T$. For a sample with class label l $(l = 1, ...L)$, the l^{th} entry of the zero-mean vector equals to a constant positive number $L - 1$.

Now, we formulate our **j**oint hete**r**ogeneous feat**u**res **l**earning (JOULE) model in the following multi-task learning framework:

$$\begin{aligned}\min_{\mathbf{W}_0, \{\mathbf{W}_i\}, \{\mathbf{\Theta}_i\}} \sum_{i=1}^{S} (&\|(\mathbf{W}_0 + \mathbf{W}_i)^T \mathbf{\Theta}_i^T \mathbf{X}_i - \mathbf{Y}\|_F^2 + \beta \|\mathbf{W}_i\|_F^2 \\ &+ \gamma \|\mathbf{X}_i - \mathbf{\Theta}_i \mathbf{\Theta}_i^T \mathbf{X}_i\|_F^2) + \alpha \|\mathbf{W}_0\|_F^2 \\ s.t. &\mathbf{\Theta}_i^T \mathbf{\Theta}_i = I, i = 1, 2, ...S\end{aligned} \tag{1}$$

where $\| \bullet \|_F$ denotes the Frobenius matrix norm. The regularization terms $\|\mathbf{W}_i\|_F^2$ and $\|\mathbf{W}_0\|_F^2$ are defined in the way such that a reliable generalization and effective closed-form solution can be obtained for our joint learning model. α and β are two parameters to control the trade-off between the shared and specific components.

Our model is casted as a least-square problem with both *prediction* (first term) and *reconstruction* loss (third term). It intends to jointly learn the common subspaces, shared and feature-specific components in a unified framework. The prediction loss item minimizes the empirical risk of each feature and thus guides our shared-specific structures learning for the purpose of better recognition. The reconstruction loss term is employed to ensure that a good reconstruction (controlled by the parameter γ) can be derived in the learned subspace using the projection matrix during optimization, which leads to a meaningful solution of the model (Eq. (1)). Here, an orthogonal constraint $\mathbf{\Theta}_i^T \mathbf{\Theta}_i = I$ is imposed on the projection matrix $\mathbf{\Theta}_i$ in order to reduce the redundancy to certain extent while preserving data information.

It is worth noting that

$$\|\mathbf{X}_i - \mathbf{\Theta}_i \mathbf{\Theta}_i^T \mathbf{X}_i\|_F^2 = \|\mathbf{X}_i\|_F^2 - \|\mathbf{\Theta}_i^T \mathbf{X}_i\|_F^2 \tag{2}$$

By substituting Eq. (2) in (1) and discarding the constant term $\|\mathbf{X}_i\|_F^2$, the function in (1) can be rewritten as:

$$\min_{\mathbf{W}_0,\{\mathbf{W}_i\},\{\mathbf{\Theta}_i\}} \sum_{i=1}^{S}(\|(\mathbf{W}_0+\mathbf{W}_i)^T\mathbf{\Theta}_i^T\mathbf{X}_i-\mathbf{Y}\|_F^2+\beta\|\mathbf{W}_i\|_F^2 - \gamma\|\mathbf{\Theta}_i^T\mathbf{X}_i\|_F^2)+\alpha\|\mathbf{W}_0\|_F^2 \tag{3}$$

3.2. Optimization

The optimization can be achieved by iterating the following three steps.

STEP 1. Fixing the coefficients $\mathbf{W}_i$ and $\mathbf{\Theta}_i$, minimize the function over $\mathbf{W}_0$:

$$\min_{\mathbf{W}_0}\sum_{i=1}^{S}\|(\mathbf{W}_0+\mathbf{W}_i)^T\mathbf{\Theta}_i^T\mathbf{X}_i-\mathbf{Y}\|_F^2+\alpha\|\mathbf{W}_0\|_F^2 \tag{4}$$

This is an unconstrained minimization problem, whose solution can be given by $\mathbf{W}_0 = (\sum_i \mathbf{\Theta}_i^T\mathbf{X}_i\mathbf{X}_i^T\mathbf{\Theta}_i + \alpha\mathbf{I})^{-1}\sum_i(\mathbf{\Theta}_i^T\mathbf{X}_i(\mathbf{Y}^T-\mathbf{X}_i^T\mathbf{\Theta}_i\mathbf{W}_i))$.

STEP 2. Fixing the coefficients $\mathbf{W}_0$ and $\mathbf{\Theta}_i$, optimize $\mathbf{W}_i$:

$$\min_{\{\mathbf{W}_i\}}\sum_{i=1}^{S}(\|(\mathbf{W}_0+\mathbf{W}_i)^T\mathbf{\Theta}_i^T\mathbf{X}_i-\mathbf{Y}\|_F^2+\beta\|\mathbf{W}_i\|_F^2)$$

The above problem can be decoupled into S independent Frobenius-regularized unconstrained least square problems:

$$\min_{\mathbf{W}_i}\|(\mathbf{W}_0+\mathbf{W}_i)^T\mathbf{\Theta}_i^T\mathbf{X}_i-\mathbf{Y}\|_F^2+\beta\|\mathbf{W}_i\|_F^2 \tag{5}$$

It is straightforward to obtain the optimal value by setting $\mathbf{W}_i = (\mathbf{\Theta}_i^T\mathbf{X}_i\mathbf{X}_i^T\mathbf{\Theta}_i+\beta\mathbf{I})^{-1}\mathbf{\Theta}_i^T\mathbf{X}_i(\mathbf{Y}^T-\mathbf{X}_i^T\mathbf{\Theta}_i\mathbf{W}_0)$.

STEP 3. Finally, we fix $\mathbf{W}_0, \mathbf{W}_i$ and optimize $\mathbf{\Theta}_i$:

$$\min_{\mathbf{\Theta}_i}\sum_{i=1}^{S}(\|(\mathbf{W}_0+\mathbf{W}_i)^T\mathbf{\Theta}_i^T\mathbf{X}_i-\mathbf{Y}\|_F^2-\gamma\|\mathbf{\Theta}_i^T\mathbf{X}_i\|_F^2)$$
$$s.t.\mathbf{\Theta}_i^T\mathbf{\Theta}_i=\mathbf{I}, i=1,2,...S$$

Note that all the $\mathbf{\Theta}_i$s in the above system are independent of each other. Hence, we turn to solve the following S independent sub-problems:

$$\min_{\mathbf{\Theta}_i}\|(\mathbf{W}_0+\mathbf{W}_i)^T\mathbf{\Theta}_i^T\mathbf{X}_i-\mathbf{Y}\|_F^2-\gamma\|\mathbf{\Theta}_i^T\mathbf{X}_i\|_F^2 \quad s.t.\mathbf{\Theta}_i^T\mathbf{\Theta}_i=\mathbf{I} \tag{6}$$

It is difficult to solve the problem (6) directly on Euclidean space due to the non-convex constraints. We optimize each sub-problem with a gradient based method on the Stiefel manifold where the approximate solution is required to satisfy the orthogonality constraint in each iteration [31]. Specifically, given the t-th step estimator of $\mathbf{\Theta}_i(t)$, we first define a skew-symmetric matrix $\nabla = \mathbf{G}\mathbf{\Theta}_i(t)^T - \mathbf{\Theta}_i(t)\mathbf{G}^T$, where $\mathbf{G}$ is the gradient of the objective function in the Euclidean space and it can be indicated by $\mathbf{G} = \mathbf{X}_i((\mathbf{W}_0+\mathbf{W}_i)^T\mathbf{\Theta}_i(t)^T\mathbf{X}_i-\mathbf{Y})^T(\mathbf{W}_i+\mathbf{W}_0)^T - 2\gamma\mathbf{X}_i\mathbf{X}_i^T\mathbf{\Theta}_i(t)$. Then the new updated point can be determined by the Grank-Nicolson-like scheme $\mathbf{\Theta}_i(t+1) = (\mathbf{I}+\frac{\tau}{2}\nabla)^{-1}(\mathbf{I}-\frac{\tau}{2}\nabla)\mathbf{\Theta}_i(t)$, where τ is the iteration step size and an optimal step size would be determined by a line search method within each iteration. We summarize our optimization for (3) in Algorithm 1.

Algorithm 1 Our method for optimizing problem (3). Note that terms $objUpdate$ and $objUpdateIn_i$ indicate the value variation of the function (3) and the $i-th$ sub-problem (6) in STEP 3, respectively.

Require:
Input: $M, \alpha, \beta, \gamma, \mathbf{Y}, \mathbf{X}_i$;
Initialization: $\mathbf{W}_0, \mathbf{W}_i \in \mathbb{R}^{M\times L}$ are random matrices, $\mathbf{\Theta}_i$ is set as the top M principal components of $\mathbf{X}_i$, $IterOut = 1$;
Ensure:
while $objUpdate \geq thr$ and $IterOut < maxIter$ **do**
  $\mathbf{W}_0 \leftarrow (\sum_i \mathbf{\Theta}_i^T\mathbf{X}_i\mathbf{X}_i^T\mathbf{\Theta}_i + \alpha\mathbf{I})^{-1}\sum_i\mathbf{\Theta}_i^T\mathbf{X}_i(\mathbf{Y}^T-\mathbf{X}_i^T\mathbf{\Theta}_i\mathbf{W}_i)$;
  $\mathbf{W}_i \leftarrow (\mathbf{\Theta}_i^T\mathbf{X}_i\mathbf{X}_i^T\mathbf{\Theta}_i+\beta\mathbf{I})^{-1}\mathbf{\Theta}_i^T\mathbf{X}_i(\mathbf{Y}^T-\mathbf{X}_i^T\mathbf{\Theta}_i\mathbf{W}_0)$, i=1,2,...,S;
  for i=1;$i\leq S$;$i++$ **do**
    while $objUpdateIn_i \geq thr$ **do**
      $\mathbf{G} \leftarrow \mathbf{X}_i((\mathbf{W}_0+\mathbf{W}_i)^T\mathbf{\Theta}_i^T\mathbf{X}_i-\mathbf{Y})(\mathbf{W}_i+\mathbf{W}_0)^T - 2\gamma\mathbf{X}_i\mathbf{X}_i^T\mathbf{\Theta}_i$
      $\nabla \leftarrow \mathbf{G}\mathbf{\Theta}_i^T - \mathbf{\Theta}_i\mathbf{G}^T$
      $\mathbf{\Theta}_i \leftarrow (\mathbf{I}+\frac{\tau}{2}\nabla)^{-1}(\mathbf{I}-\frac{\tau}{2}\nabla)\mathbf{\Theta}_i$;
    end while
  end for
  $IterOut++$;
end while
return $\mathbf{W}_0, \mathbf{W}_i, \mathbf{\Theta}_i$

3.3. Inference

Given the model parameters $\mathbf{W}_0, \mathbf{W}_i$ and $\mathbf{\Theta}_i$, the inference is to find the best activity label for a new sample with heterogeneous features $\mathbf{x}_i, i = 1, 2....S$. We first define two confidence vectors to encode the shared and specified components of $\mathbf{x}_i$ as

$$\mathbf{C}_{shared}^i = \mathbf{W}_0^T\mathbf{\Theta}_i^T\mathbf{x}_i \in \mathbb{R}^L$$
$$\mathbf{C}_{specified}^i = \mathbf{W}_i^T\mathbf{\Theta}_i^T\mathbf{x}_i \in \mathbb{R}^L$$

Based on the formulated confidence vectors, we then developed two inference models.

JOULE-Score. In this inference model, the final assignment is derived by finding the label with the maximum score

$$l^* = arg\max_l \sum_{i=1,2,...S}[\mathbf{C}_{shared}^i(l) + \mathbf{C}_{specified}^i(l)]$$

We denote this model as "JOULE-Score". This is a naive but effective inference model.

JOULE-SVM. Inspired by the construction of augmented feature in [12], we treat all the shared and specific confidence vectors as higher-level augmented features and concatenate them together to form our final representation. To speed up our testing, a linear SVM classifier is employed to make the final decision. This reference model is referred as "JOULE-SVM".

4. Experiments

We evaluated our methods on two benchmark 3D activity datasets and a newly collected human-object interaction dataset. In the following, we first briefly introduce the implementation details and then describe the experiments and results.

4.1. Implementation Details

The model parameters α, β, γ are fixed as $10^{-1}, 10^{-1}$ and 1, respectively through all our experiments. The dimensionality M of the subspace is specified empirically for each dataset. Intuitively, it is suggested to be smaller than the number of training samples. We investigate its effect in detail in Section 4.5. When computing DCP and DDP features, one image patch of size 60×60 is extracted around each joint position in a trajectory, which is large enough to capture the context cues. A set of image patches (RGB or Depth) are extracted per trajectory. For computational efficiency, all the image patches are then resized to 32×32 and the cell size of HOG is set to 8.

4.2. MSR Daily Activity Dataset

We test the proposed methods on a 3D activity set, MSR Daily Activity dataset [29], which has become a de facto standard set for studying 3D human activities. It contains 320 video clips of 16 different activities (drinking eat, reading book, etc) performed by 10 subjects in 2 different poses *sitting* and *standing*. Most of the activities involve human-object interactions (see Table 3). We follow the same experimental settings as other related works, where half of the subjects are used for training and the rest for testing.

To evaluate our proposed methods, JOULE-Score and JOULE-SVM, we compare them with a baseline implementation that fuses different features together with a standard SVM classifier. We denote this baseline as 'X+SVM', where 'X' denotes the employed feature type. It can be 'DCP', 'DS', or 'DDP', or even their combination 'DS+DCP+DDP' by feature concatenation. We also report the result obtained by an implementation using the framework developed in MTDA [39], which is denoted as 'DS+DCP+DDP+MTDA'. In particular, we present the recently reported results of other 10 different methods for comparison. The dimensionality M for JOULE based methods is set to 40.

	Method	Accuracy
Reported Results	Dynamic Temporal Warping [19]	54
	3D Joints and LOP Fourier [29]	78
	HON4D [22]	80.00
	SSFF [23]	81.9
	Deep Model (RGGP) [15]	85.6
	Actionlet Ensemble [29]	85.75
	Super Normal [36]	86.25
	DCSF+Joint [32]	88.2
	Group Sparsity [17]	95
	Range Sample [16]	**95.6**
Our Results	DS+SVM	71.25
	DCP+SVM	86.25
	DDP+SVM	83.13
	DS+DCP+DDP+SVM	90
	DS+DCP+DDP+MTDA [39]	90.62
	DS+DCP+DDP+JOULE-Score	93.13
	DS+DCP+DDP+JOULE-SVM	**95**

Table 1. Comparison on the MSR Daily Activity dataset.

Results. Table 1 shows the results and comparison. Our method obtains an accuracy of 95%, which outperforms most of the latest reported results and is comparable with the state-of-the-art [16]. Compared with closely related work focusing on feature fusion using deep model [15] and structured sparse model [23], our model outperforms both of them by a considerable margin (more than 9.4%), which implies our method is superior to other RGB-D activity fusion systems. Compared with our baselines, our JOULE-SVM model obtains the best performance as expected. Especially, compared with 'DS+DCP+DDP+SVM', the performance gain (e.g., 95% vs. 90%) of both our JOULE-Score and JOULE-SVM models demonstrates the benefits of the shared and specific components modelling. Our JOULE-SVM outperforms MTDA considerably by 4.4% using exactly the same set of features, which demonstrates the effectiveness of the proposed joint learning framework.

The confusion matrix of the results by our JOULE-SVM model is shown in Figure 3. It can be seen that our model achieves perfect classification results on 10 classes. The larger error is due to the mis-classification of the activity of *writing on a paper* as *reading book*, which may be largely attributed to high similarity between the object and activity contexts in these two activities.

4.3. Cornell Activity Dataset 60 (CAD 60)

This public dataset consists of 68 video clips captured by Microsoft Kinect device [25]. Four actors were asked to perform 13 specific activities (*still*, *talking on the phone*, and etc.) and one random activity in 5 different environments: office, kitchen, bedroom, bath room, and living room. We follow the same experimental setting as in [29] by adopting the leave-one-person-out cross validation per environment, which ensures that person participating in the

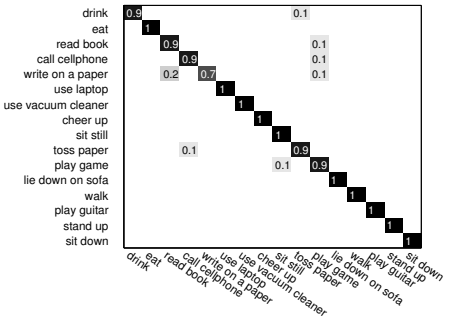


Figure 3. Confusion matrix of JOULE-SVM on MSR Daily dataset.

training cannot be seen in the testing. The final accuracy was calculated by averaging the accuracies of all the possible splits (totally 20 in this set).

Our methods are compared with the results reported in the state-of-the-art [29]. We also ran the released code of HON4D [1] on this set. Since there is no default parameter settings suggested by the author on this set, we report the best results by varying their parameters in a wide range. Similar to MSR Daily set, we also highlight the benefits of using JOULE model by comparing with the baselines 'X+SVM' and 'DS+DCP+DDP+MTDA'. The dimensionality M of the subspace is set as 4 for this dataset.

Results. The results and comparison are shown in Table 2. Our method achieves an accuracy of 84.1%, which significantly outperforms the state-of-the-art result [29] by a large margin (9.4%). It is worth noting that all our baseline implementations including the simple combination of our DCP, DDP or DS features with a SVM classifier outperform the state-of-the-art by more than 1.1%, which proves that our feature is superior to that developed in [29]. As expected, fusing them together would further improve the performance. Especially, by considering the shared and specific components, our model (JOULE-SVM) obtains a gain of 9.1% compared with the fusion methods using standard SVM classifier without explicitly modelling shared and specific components (84.1% vs. 75%). Our JOULE-SVM works better than MTDA on the CAD 60 set with a smaller performance gain than on the MSR Daily set.

The confusion matrix of the results by our JOULE-SVM model is presented in Figure 4. It can be seen that our model can distinguish well the five activities of rinsing mouth with water, wearing contact lenses, cooking(chopping), working on computer and random activities, which demonstrates that our model can effectively capture the interactions between human and the manipulated object. It can also be observed that the activities of talking on couch and relaxing on couch are often confused by our model, mainly due to the inaccurate human skeletons captured by the Kinect camera.

[1] http://www.cs.ucf.edu/~oreifej/HON4D.html

	Method	Accuracy
Reported Results	STIP [41]	62.5
	Order Sparse Coding [20]	65.3
	Object Affordance [10]	71.4
	HON4D [22]	72.7
	Actionlet Ensemble [29]	74.7
Our Results	DS+SVM	76.5
	DLP+SVM	75.8
	DDP+SVM	77.9
	DS+DCP+DDP+SVM	75
	DS+DCP+DDP+JOULE-Score	81.1
	DS+DCP+DDP+MTDA [39]	82.6
	DS+DCP+DDP+JOULE-SVM	**84.1**

Table 2. Comparison on the CAD 60 dataset.

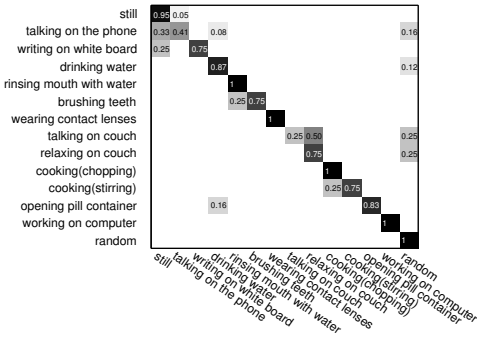


Figure 4. Confusion matrix of JOULE-SVM on CAD 60 set.

4.4. SYSU 3D Human-Object Interaction Dataset

Dataset Description. We collect a new RGB-D activity dataset focusing on human-object interactions to further evaluate our method. We name this as *SYSU 3D Human-Object Interaction* (HOI) dataset. For constructing this set, 40 subjects were asked to perform 12 different activities. For each activity, each participants manipulate one of the six different objects: phone, chair, bag, wallet, mop and besom. Therefore, there are totally 480 video clips collected in this set. For each video clip, the corresponding RGB frames, depth sequence and skeleton data are captured by a Kinect camera. Sample activities are shown in Figure 6. We highlight the differences between our 3D HOI set and relevant existing sets in Table 3. Compared to those datasets, our new dataset presents new challenges: 1) the involved motions and the manipulated objects' appearance are highly similar between some activities; 2) the number of participants is at least four times larger than that of existing ones. To the best of our knowledge, this is the most complete set to date for studying 3D activities in terms of the number of subjects.

Evaluation Protocol. We test our methods in two different settings. For the first setting (setting-1), for each activity class, we select half of the samples for training and the rest for testing. For the second setting (setting-2), sequences performed by half of the subjects are used to learn model parameters and the rest for testing, where there is no over-

DataSet	Data	Cla. No.	Sub. No.	Vid. No.	HOI Ra.
CAD 60 [25]	RGB-D	14	4	68	85.7%
MSRDaily [29]	RGB-D	16	10	320	87.5%
MSRAction [13]	Depth	20	10	567	$\leq 70\%$
Multiview [30]	RGB-D	8	8	3815	100%
3D HOI	RGB-D	12	40	480	100%

Table 3. Comparison of 3D HOI dataset with relevant datasets. Cla. denotes *class*, and Sub. for *subject*, Vid for *video*, HOI Ra. for HOI *ratio* among the dataset.

Method	Mean Acc±std (%)	
	setting-1	setting-2
HON4D [22]	73.39 (±2.59)	79.22 (±2.36)
DS+SVM	74.41 (±2.29)	75.53 (±3.08)
DCP+SVM	70.57 (±2.31)	77.32 (±2.48)
DDP+SVM	71.82 (±2.24)	78.29 (±2.44)
DS+DCP+DDP+SVM	77.34 (±2.53)	82.78 (±2.83)
DS+DCP+DDP+MTDA [39]	79.19 (±4.27)	84.21 (±2.19)
DS+DCP+DDP+JOULE-Score	79.22 (±1.78)	84.24 (±2.23)
DS+DCP+DDP+JOULE-SVM	**79.63 (±2.13)**	**84.89 (±2.29)**

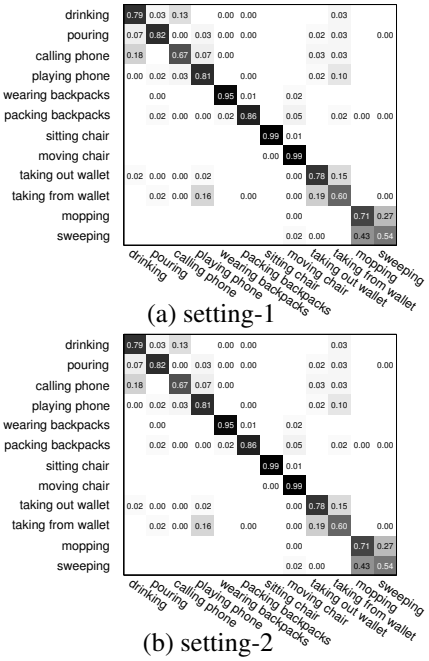
Table 4. Comparison on SYSU 3D HOI dataset.

lap of subjects between the training and test set. This is a *cross-subject* setting. For each setting, we report the mean accuracy and standard deviation of the results over 30 random splits.

Baselines. We test the HON4D method by running its released code on this dataset, and report its best performance over a wide range of parameters. Similar to that in MSRDaily and CAD60 set, the baselines "X+SVM" and 'DS+DCP+DDP+MTDA' are also compared to show the effectiveness of our joint learning models (JOULE-Score and JOULE-SVM). In total, we report a comprehensive set of results of up to *eight* different implementations on this set.

Results. We empirically observed that $M = 30$ is sufficient for obtaining a good performance for our model. Table 4 reported the results. Again, using the proposed JOULE model to fuse different heterogeneous features is always beneficial in all settings. The accuracies in setting-2 are higher than that of setting-1 without considering cross-subject split. This is because the prediction could be biased by appearance when activities with similar motion and object context (e.g. mopping vs. sweeping) performed by the same subject are contained in both training and test sets, which may occur in the setting-1. The performances of JOULE-SVM and MTDA are comparable with JOULE-SVM performing slightly better. It was noted that the performance gap between our models and the baselines is smaller (e.g., 84.9% vs. 82.8%) than that of other two datasets. This somehow indicates the new dataset is quite challenging for feature fusion.

By examining the confusion matrices of our JOULE-SVM model in Figure 5, we observed that our model often confuses the activities of mopping with sweeping in both settings, which is mainly due to similar motions and object



(a) setting-1

(b) setting-2

Figure 5. Confusion tables of JOULE-SVM on SYSU 3D HOI set.

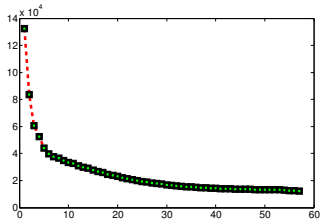


Figure 7. Illustration of the convergence of our algorithm.

appearance in the two interactions. In addition, the activities of taking from wallet share similar motions with activities of playing phone and taking out wallet, which are occasionally misidentified as playing phone or taking out wallet.

4.5. Analysis and Discussion

Convergence. Our method converges to a minimum after a limited number of iterations. We empirically observed that 20 iterations (outer iterations i.e. term $IterOut$ in Algorithm 1) are sufficient for obtaining a reliable solution in all of our experiments. See Figure 7 for an example illustrating the convergence of our method on the MSR Daily activity dataset, where the objective function value of each step was recorded during each iteration. Excluding the time for computing the features, one round training of our algorithm takes about 1.26 minutes per training sample on a machine of 8-core (a MATLAB worker pool with maximum of 4 processes). Our testing is pretty fast, and takes about 0.5 second per test sample.

Effect of dimensionality M. We investigate the effect of the dimensionality M of the subspace. Figure 8 shows the

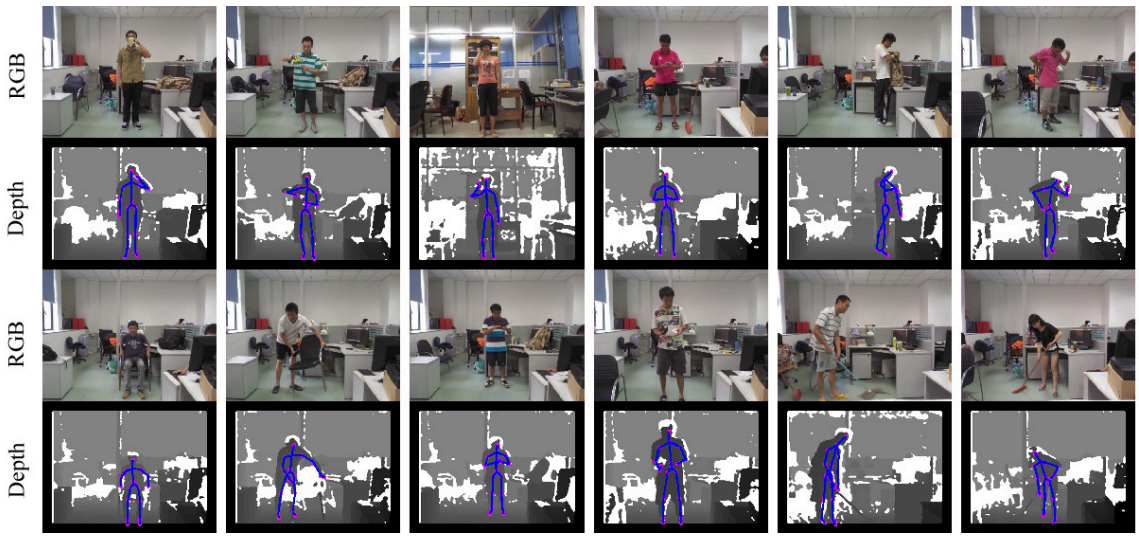
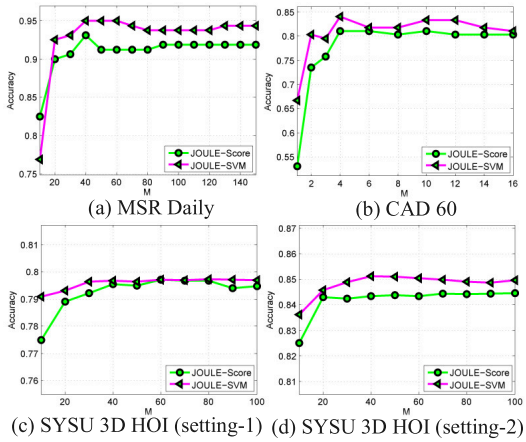


Figure 6. Snapshots of activities in SYSU 3D HOI set, one sample per class. The rows headed with *RGB* show the samples in RGB channel and the rows underneath headed with *Depth* show the corresponding depth channel superimposed with skeleton data. Best viewed in color.



(a) MSR Daily (b) CAD 60

(c) SYSU 3D HOI (setting-1) (d) SYSU 3D HOI (setting-2)

Figure 8. Effects of parameter M on the system performance.

performances of our methods including JOULE-Score and JOULE-SVM with different values of M. Generally, very small value of M leads to a worse performance, as the lower dimensionality of the subspace, the less representative for the original features. When M becomes larger (typically larger than a value about $\frac{1}{6} \sim \frac{1}{4}$ of the number of training samples), the performances start to remain stable, which means our algorithm is not sensitive to the value of M in a reasonable range. JOULE-SVM consistently outperforms JOULE-Score in most of the cases.

Effect of TPF on Gradient Signal. Table 5 shows the results of our model with and without temporal Fourier features computed from the gradient signal on all of the three datasets. It can be seen that, while the improvement on the 3D HOI dataset is relatively mild, TPF features on gradient consistently improve the results in all of the cases, with the biggest gain (7.6%) achieved on the CAD60 dataset. This indicates that the proposed extension of TPF features to the gradient signal is promising and effective.

	MSRD	CAD60	3DHOI(s-1)	3DHOI(s-2)
Full	95(93.13)	84.1(81.1)	79.63(79.22)	84.89(84.24)
NoG.	91.25(88.75)	76.5(75)	78.83(78.59)	83.63(83.12)

Table 5. Accuracy (%) of our methods with and without TPF on gradient (denoted by Full and NoG respectively). s-1 denotes setting-1 and s-2 for setting-2 applied on the 3D HOI dataset. Accuracies outside (inside) the parentheses are achieved by JOULE-SVM (JOULE-Score).

5. Conclusion and Future Work

We have proposed a new method called **j**oint het-er**o**geneous feat**u**res **le**arning (JOULE) model to fuse the heterogeneous features for RGB-D activity recognition. State-of-the-art results are achieved on three 3D activity sets, which demonstrated the effectiveness of the proposed method. One direction of our future work is to explore the application of kernel learning for our shared and feature-specific components modeling.

Acknowledgment

This work was supported in part by the National Natural Science of Foundation of China (Grant No. 61472456, 61173084, U1135001), the 12th Five-year Plan China Science and Technology Supporting Programme under Grant 2012BAK16B06, the Guangzhou Pearl River Science and Technology Rising Star Project under Grant 2013J2200068, the Guangdong Natural Science Funds for Distinguished Young Scholar under Grant S2013050014265, and in part by RSE-NSFC joint project (RSE Reference: 443570/NNS/INT).

References

[1] Y. Amit, M. Fink, N. Srebro, and S. Ullman. Uncovering shared structures in multiclass classification. In *ICML*, 2007. 2

[2] R. K. Ando and T. Zhang. A framework for learning predictive structures from multiple tasks and unlabeled data. *JMLR*, (6):1817–1853, 2005. 2

[3] L. Cao, J. Luo, F. Liang, and T. S. Huang. Heterogeneous feature machines for visual recognition. In *ICCV 2009*. 2

[4] A. A. Chaaraoui, J. R. Padilla-López, and F. Flórez-Revuelta. Fusion of skeletal and silhouette-based features for human action recognition with rgb-d devices. In *ICCVW*, 2013. 1

[5] J. Chen, L. Tang, J. Liu, and J. Ye. A convex formulation for learning a shared predictive structure from multiple tasks. *TPAMI*, 35(5):1025–1038, 2013. 2

[6] P. Dollár. Piotr's Computer Vision Matlab Toolbox (PMT). http://vision.ucsd.edu/~pdollar/toolbox/doc/index.html. 3

[7] S. Han, X. Liao, and L. Carin. Cross-domain multitask learning with latent probit models. In *ICML*, 2012. 2

[8] J. Hu, W. Zheng, J. Lai, S. Gong, and T. Xiang. Exemplar-based recognition of human-object interactions. *TCSVT*, PP(99):1–1, 2015. 3

[9] M. Hussein, M. Torki, M. Gowayyed, and M. El-Saban. Human action recognition using a temporal hierarchy of covariance descriptors on 3d joint locations. In *IJCAI*, 2013. 1

[10] H. S. Koppula, R. Gupta, and A. Saxena. Learning human activities and object affordances from rgb-d videos. *IJRR*, 32(8):951–970, 2013. 1, 6

[11] J. Lei, X. Ren, and D. Fox. Fine-grained kitchen activity recognition using rgb-d. In *Ubicomp*, 2012. 1

[12] W. Li, L. Duan, D. Xu, and I. W. Tsang. Learning with augmented features for supervised and semi-supervised heterogeneous domain adaptation. *TPAMI*, 36(6):1134–1148, 2014. 5

[13] W. Li, Z. Zhang, and Z. Liu. Action recognition based on a bag of 3d points. In *CVPRW*, 2010. 1, 7

[14] I. Lillo, A. Soto, and J. C. Niebles. Discriminative hierarchical modeling of spatio-temporally composable human activities. In *CVPR*, 2014. 1

[15] L. Liu and L. Shao. Learning discriminative representations from rgb-d video data. In *IJCAI*, 2013. 1, 5

[16] C. Lu, J. Jia, and C.-K. Tang. Range-sample depth feature for action recognition. In *CVPR*, 2014. 1, 5

[17] J. Luo, W. Wang, and H. Qi. Group sparsity and geometry constrained dictionary learning for action recognition from depth maps. In *ICCV*, 2013. 1, 5

[18] F. Lv and R. Nevatia. Recognition and segmentation of 3-d human action using hmm and multi-class adaboost. In *ECCV*. 2006. 1

[19] M. Müller and T. Röder. Motion templates for automatic classification and retrieval of motion capture data. In *SCA*, pages 137–146, 2006. 5

[20] B. Ni, P. Moulin, and S. Yan. Order-preserving sparse coding for sequence classification. In *ECCV*. 2012. 6

[21] F. Ofli, R. Chaudhry, G. Kurillo, R. Vidal, and R. Bajcsy. Sequence of the most informative joints (smij): A new representation for human skeletal action recognition. *JVCIR*, 25(1):24–38, 2014. 1

[22] O. Oreifej and Z. Liu. Hon4d: Histogram of oriented 4d normals for activity recognition from depth sequences. In *CVPR*, 2013. 1, 5, 6, 7

[23] A. Shahroudy, G. Wang, and T.-T. Ng. Multi-modal feature fusion for action recognition in rgb-d sequences. In *ISCCSP*, pages 1–4, May 2014. 1, 5

[24] J. Shotton, T. Sharp, A. Kipman, A. Fitzgibbon, M. Finocchio, A. Blake, M. Cook, and R. Moore. Real-time human pose recognition in parts from single depth images. In *CVPR*, 2011. 1, 3

[25] J. Sung, C. Ponce, B. Selman, and A. Saxena. Human activity detection from rgbd images. *PAIR*, 64, 2011. 5, 7

[26] A. Torralba, K. Murphy, and W. Freeman. Sharing visual features for multiclass and multiview object detection. *TPAMI*, 29(5):854–869, 2007. 2

[27] R. Vemulapalli, F. Arrate, and R. Chellappa. Human action recognition by representing 3d skeletons as points in a lie group. In *CVPR*, 2013. 1

[28] J. Wang, Z. Liu, J. Chorowski, Z. Chen, and Y. Wu. Robust 3d action recognition with random occupancy patterns. In *ECCV*. 2012. 1

[29] J. Wang, Z. Liu, Y. Wu, and J. Yuan. Learning actionlet ensemble for 3d human action recognition. *TPAMI*, 36(5):914–927, 2014. 1, 2, 3, 5, 6, 7

[30] P. Wei, Y. Zhao, N. Zheng, and S.-C. Zhu. Modeling 4d human-object interactions for event and object recognition. In *ICCV*, 2013. 1, 7

[31] Z. Wen and W. Yin. A feasible method for optimization with orthogonality constraints. *MP*, 142(1-2):397–434, 2013. 4

[32] L. Xia and J. Aggarwal. Spatio-temporal depth cuboid similarity feature for activity recognition using depth camera. In *CVPR*, 2013. 5

[33] L. Xia, C.-C. Chen, and J. Aggarwal. View invariant human action recognition using histograms of 3d joints. In *CVPRW*, 2012. 1

[34] X. Yang, S. Kim, and E. P. Xing. Heterogeneous multitask learning with joint sparsity constraints. In *NIPS*, 2009. 2

[35] X. Yang and Y. Tian. Eigenjoints-based action recognition using naive-bayes-nearest-neighbor. In *CVPRW*, 2012. 1

[36] X. Yang and Y. Tian. Super normal vector for activity recognition using depth sequences. In *CVPR*, 2014. 1, 5

[37] B. Yao and L. Fei-Fei. Action recognition with exemplar based 2.5d graph matching. In *ECCV*, 2012. 3

[38] Y. Zhang and D. Yeung. A convex formulation for learning task relationships in multi-task learning. In *UAI*, 2010. 2

[39] Y. Zhang and D. Yeung. Multi-task learning in heterogeneous feature spaces. In *AAAI*, 2011. 2, 5, 6, 7

[40] Y. Zhu, W. Chen, and G. Guo. Fusing spatiotemporal features and joints for 3d action recognition. In *CVPRW*, 2013. 1

[41] Y. Zhu, W. Chen, and G. Guo. Evaluating spatiotemporal interest point features for depth-based action recognition. *IVC*, 2014. 1, 6