

Max-Margin Heterogeneous Information Machine for RGB-D Action Recognition

Yu Kong¹  · Yun Fu²

Received: 12 February 2016 / Accepted: 10 December 2016 / Published online: 21 January 2017
© Springer Science+Business Media New York 2017

Abstract We propose a novel approach, max-margin heterogeneous information machine (MMHIM), for human action recognition from RGB-D videos. MMHIM fuses heterogeneous RGB visual features and depth features, and learns effective action classifiers using the fused features. Rich heterogeneous visual and depth data are effectively compressed and projected to a learned shared space and independent private spaces, in order to reduce noise and capture useful information for recognition. Knowledge from various sources can then be shared with others in the learned space to learn cross-modal features. This guides the discovery of valuable information for recognition. To capture complex spatiotemporal structural relationships in visual and depth features, we represent both RGB and depth data in a matrix form. We formulate the recognition task as a low-rank bilinear model composed of row and column parameter matrices. The rank of the model parameter is minimized to build a low-rank classifier, which is beneficial for improving the generalization power. We also extend MMHIM to a structured prediction model that is capable of making structured outputs. Extensive experiments on a new RGB-D action dataset and two other public RGB-D action datasets show that our approaches achieve state-of-the-art results. Promising results

are also shown if RGB or depth data are missing in training or testing procedure.

Keywords Action recognition · RGB-D videos · Heterogeneous data · Feature learning

1 Introduction

Understanding visual data captured by depth sensors such as Kinect (Shotton et al. 2013) has been receiving increasing interests in the computer vision community thanks to the recent advent of cost-effective depth sensors. In addition to RGB visual data captured by conventional RGB cameras, depth data encode rich 3D structural information of the entire scene, which is one of its key benefits. Such important information has shown to be helpful in reducing background noise, and thus has already been widely leveraged in pose estimation (Xu and Cheng 2013), layout estimation (Zhang et al. 2013), tracking (Zhou et al. 2015), image understanding (Wang et al. 2015), surface recovery (El et al. 2015), and action recognition (Kong and Fu 2015).

Action recognition from easy-to-use and low-cost depth sensors, such as Kinect sensors, attempts to predict the action labels from RGB-D videos. Thanks to the extra depth data, background noise that is frequently seen in action videos can be markedly reduced, thereby boosting the classification performance. Previous work Shotton et al. (2013), Oreifej and Liu (2013), Wang et al. (2012b) and Hadfield and Bowden (2013) showed that effective usage of 3D structural information facilitates recognition tasks as it simplifies intra-class motion variations and reduces cluttered background noise. Plenty of action descriptors specifically designed for depth action videos have been proposed, for example, action graph (Li et al. 2010), histogram of oriented 4D normals (Oreifej

Communicated by M. Hebert.

✉ Yu Kong
yukong@ece.neu.edu
Yun Fu
yunfu@ece.neu.edu

¹ Department of ECE, Northeastern University, Boston, MA, USA

² Department of ECE and College of CIS, Northeastern University, Boston, MA, USA

and Liu 2013), super normal vector (Yang et al. 2014), and depth spatiotemporal interest points (Xia and Aggarwal 2013; Hadfield and Bowden 2013).

Despite their effectiveness, those methods are limited to the scenario where depth data must be available. Depth data-based methods developed in Yang et al. (2014), Oreifej and Liu (2013), Xia and Aggarwal (2013) and Hadfield and Bowden (2013) would fail if depth data are not available due to the failure of depth sensors in RGB-D sensing devices. In addition, a working depth sensor may fail to compute a depth measurement due to fundamental physical limitations, which consequently causes missing depth data problem. For example, a human subject is too close to or too far from the depth sensor or interacting objects have reflective surfaces. Furthermore, depth data normally contain spatiotemporal discontinuous regions. These regions make the depth data very noisy, and consequently hinder the application of feature extraction methods such as surface normal (Yang et al. 2014; Oreifej and Liu 2013) and spatiotemporal interest points (Xia and Aggarwal 2013; Hadfield and Bowden 2013) in these regions. If the discontinuous regions unfortunately appear in the body parts that were supposed to provide discriminative cues, such as arms or legs, the recognition performance will be undoubtedly degraded where depth information is used as a the only cue.

Visual data and depth data can be complementary to each other. Recent work El et al. (2015), Jia et al. (2014) and Wang et al. (2015) has demonstrated that the fusion of visual data and depth data can notably improve the performance. It was also shown in Jia et al. (2014) and Kong and Fu (2015) that implicit correlations between visual and depth data can be learned to handle the case where one of them is unavailable. Moreover, RGB data are robust with no discontinuities. Numerous feature descriptors (e.g. gradient and optical flow) can be extracted from RGB data, providing abundant and robust features for recognition tasks. Furthermore, human bodies consist of multiple structural objects, and thus motions of human body parts are highly correlated. Existing work for action recognition from depth sequences (Yang et al. 2014; Oreifej and Liu 2013) attempted to capture spatiotemporal correlation information of body part movements by aggregating features from neighborhoods. However, the information would unfortunately be collapsed (Tenenbaum and Freeman 2000) if co-occurrence features are concatenated into a high dimensional vector and then linearly projected onto a subspace.

In this paper, we propose a novel max-margin heterogeneous information machine (MMHIM) for action recognition from RGB-D sequences. MMHIM treats visual and depth data as two modalities, and learns both features shared between them and private features for classification. More specifically, we project the original features of the two modalities onto a shared space, and learn cross-modal features

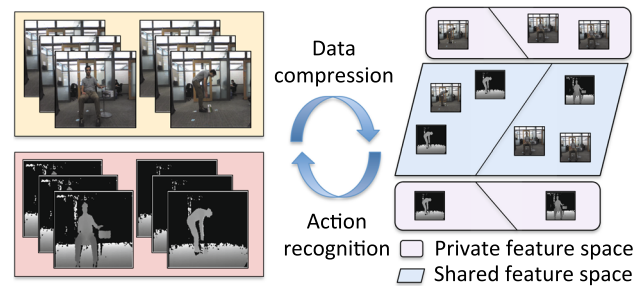


Fig. 1 MMHIM projects and compresses both RGB visual features and depth features to a learned shared feature space and modality-specific private feature spaces. Features in these spaces are all used for learning classification boundaries. These two steps iterate until convergence

shared between them for classification in order to effectively capture cross-modal knowledge. The learned cross-modal features inherit the characteristics of both RGB and depth data that capture motion, 3D structural, and spatiotemporal relationship information. In addition, private features are learned by projecting the original features to modality-specific spaces. These private features capture unique and intrinsic information of a modality, for example, edge cue in visual modality and distance cue in depth data. The use of both shared cross-modality features and private features allows us to leverage all discriminative information for classification. These two types of features are learned through data compression and noise “filtering” during the projection procedure, and are jointly optimized with the learning of the action classifiers. This automatically discovers compact yet discriminative features for the classifiers, and thus improves recognition performance (see Fig. 1). A structured prediction model is also proposed in this work, which allows us to model the agreement between low-level heterogeneous features and high-level structured outputs. We show in the experiment that the learned features are expressive and discriminative for differentiating action categories, even if one modality is missing in training or testing.

We represent both visual and depth features in a matrix form, which naturally encodes spatiotemporal structural relationships. Even though feature matrices are projected onto a low-dimensional space, the structural information of body parts is conserved and motion information is compressed and denoised. This overcomes the aforementioned problem of the collapsed information in feature vectors.

The recognition problem is formulated in a low-rank bilinear framework, particularly designed for feature representations in a matrix form. The proposed model learns feature projection matrices and classification parameter matrices, which operate as feature weighting in both rows and columns, respectively. The projection matrices are optimized to map original heterogeneous visual and depth features onto a

shared feature space and private feature spaces. The shared space is the optimal space for building robust and effective cross-modal features for recognition; while the features in a private space inherit unique information of the corresponding modality. An information measure is incorporated in the learning of projection matrices to help reduce noise in the feature projection procedure. Classification is performed using both the learned cross-modal features and private features. The rank of the model is minimized to increase the generalization power and decrease computational cost (Wolf et al. 2007).

We propose an efficient algorithm to optimize MMHIM. Without approximations nor hard constraint on the rank of the parameter matrices, we present a regularized risk minimization problem that produces low-rank projection matrices and action classifiers by minimizing the Frobenius norm of the parameter matrices. This allows us to use existing efficient SVM solvers. The learning problem is iteratively solved with a bundle method (Teo et al. 2007; Do and Artieres 2009) being the solver for the inner optimization problem.

The main contribution of this work is the MMHIM, a novel formalism for RGB-D action recognition. With inputs of feature matrices rather than vectors, MMHIM keeps inherent spatiotemporal structural information within features, which plays a key role in recognition. In addition, MMHIM learns a shared space for fusing heterogeneous data (RGB and depth data in this work), where knowledge can be shared between them. MMHIM directly minimizes the rank of parameter matrices, and produces compact yet expressive cross-modal features through the use of information measure. MMHIM is also able to handle structured prediction problem. An efficient solver is developed for MMHIM in this work. We show that MMHIM achieves superior performance over state-of-the-art methods.

1.1 Overview of Our Approach

We study the use of both visual and depth data for action recognition from RGB-D videos. Note that a RGB-D has two channels, a RGB color video that is similar to a conventional color video used in action recognition, and an extra depth video that encodes distance information from the human subject to the camera.

The flowchart of the training procedure in our approach is illustrated in Fig. 2. Our approach consists of the following four main steps:

1. Preprocessing. Our approach uniformly samples a fixed number K of frames for all videos in a dataset. This procedure makes all the videos have equal length.
2. Low-level feature extraction. Given a RGB-D video sample (contains a color video and a depth video), we

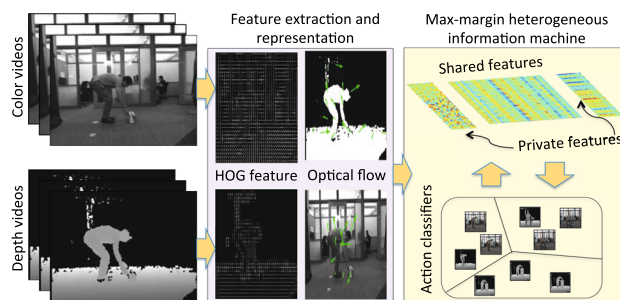


Fig. 2 Flowchart of our approach in training. Please refer to Sect. 1.1 for details

compute the gradient and the optical flow over both the color video and the depth video.

3. Feature representation. The histograms of oriented gradients (HOG) and histograms of optical flow (HOF) descriptors are adopted to represent the gradient and optical flow, respectively. Consequently, each channel (color video or depth video) of a RGB-D video will generate two features, i.e., HOG features and HOF features.
4. Model learning. The proposed MMHIM takes the low-level HOG and HOF features extracted from color and depth videos as inputs. Compact yet discriminative features are then summarized by MMHIM, and used for classification. Note that MMHIM jointly learns discriminative features and action classifiers at the same time.

Given a testing RGB-D video, our approach first samples K frames from each channel of the testing sample. Then HOG and HOF features are extracted over both the preprocessed color video and the depth video. The HOG and HOF action descriptors are fed into the trained MMHIM to compute the shared and private features, and finally the action label is predicted based on the computed features.

2 Related Work

2.1 Action Recognition from RGB-D Videos

Previous action recognition approaches mainly focus on RGB action videos (Kong et al. 2014; Raptis and Sigal 2013; Tang et al. 2012; Ji et al. 2013). These studies used low-level interest point features (Tang et al. 2012), mid-level semantic features (Kong et al. 2014) or human pose (Raptis and Sigal 2013), or learned features using deep learning techniques (Ji et al. 2013). However, misclassification exists due to large intra-class motion and pose variations.

Thanks to the advent of low-cost Kinect sensors (Shotton et al. 2013), much effort has been devoted to object recognition (Chen et al. 2014; Bo et al. 2011; Wang et al. 2015) and action recognition (Li et al. 2010; Oreifej and

Liu 2013; Yang et al. 2014; Hadfield and Bowden 2013; Wang et al. 2012a, b) from depth images. A group of RGB-D or depth action video datasets were introduced in Ni et al. (2011), Wang et al. (2012b), Oreifej and Liu (2013), Hadfield and Bowden (2013) and Offi et al. (2013) such as RGBD-HuDaAct dataset, MSR Pair Actions dataset, and Hollywood 3D dataset. These studies showed that depth data capture 3D structural information, which helps reduce background noise and intra-class variations. Effective features have been proposed for the recognition task using depth data, such as histogram of oriented 4D normals (Oreifej and Liu 2013; Yang et al. 2014) and depth spatiotemporal interest points (Xia and Aggarwal 2013; Hadfield and Bowden 2013). Features from depth sequences can be encoded by Luo et al. (2013), or be used to build actionlets (Wang et al. 2012b) for recognition. The work in Sung et al. (2012) and Koppula and Saxena (2013) built layered action graph structures to model actions and subactions in a RGB-D video. Recent work Liu and Shao (2013) also showed that features of RGB-D data can be learned using deep learning techniques.

The methods in Li et al. (2010), Oreifej and Liu (2013), Yang et al. (2014), Hadfield and Bowden (2013), Wang et al. (2012a), and Luo et al. (2013) only use depth data, and thus would fail if depth data were missing. In contrast, our method uses both RGB and depth data, and can handle the case if one modality is missing. Even though existing work Hu et al. (2015), Jia et al. (2014), Lin et al. (2014), Liu and Shao (2013) and Wang et al. (2012b) used RGB and depth modalities, they only learned features shared between the two modalities and did not learn modality-specific or private features. Private features capture unique information of one modality and provide extra discriminative information for classification. In addition, the methods in Hu et al. (2015), Lin et al. (2014), Liu and Shao (2013) and Wang et al. (2012b) were not developed for dealing with missing modality problem and their performance in missing modality scenario is unknown. Moreover, they used features in a vector form, in which spatiotemporal structures would not be well represented (Tenenbaum and Freeman 2000; Kobayashi 2014). In this work, we use features in a matrix form (Pirsiavash et al. 2009), which naturally captures both spatiotemporal structural information and motion information. Our experiments show that features in a matrix format significantly improve the performance even though the rank of the parameter matrices in MMHIM is constrained to be 1.

An unsupervised setting was considered in Wu et al. (2015). In their work, long-range action relations such as the occurrence of put-milk-back-to-fridge and fetch-milk-from-fridge is modeled in RGB-D temporal sequences. Heterogeneous feature learning was also investigated in Hu et al. (2015). Their method projects various types of features including skeleton features and local HOG features into a shared feature space, and then uses SVM as the classifiers.

The projection matrices are learned by minimizing the reconstruction loss. Different from this work, our approach jointly learns heterogeneous features and action classifiers. The projection matrices in our work are learned by minimizing the noise after projection and classification error using the projected features. The solution to the recognition task leverages auxiliary databases was studied in Jia et al. (2014) and Lin et al. (2014). Their methods assume actions can be reconstructed by entries in the auxiliary databases. Instead of using an auxiliary database to compute action representations, our method uses the information from the classifiers to guide the learning of discriminative action representations. This will learn the features that are optimized for classification. An efficient binary range-sample feature for depth data was proposed in Lu et al. (2014). This new type of depth feature has shown to be invariant to possible changes in scale, viewpoint, and background, and it is fast due to the binary property.

2.2 Action Recognition from Color Videos

In recent studies, human actions were popularly represented by local salient features detected by spatio-temporal interest points (Dollar et al. 2005; Schödl et al. 2004; Laptev 2005; Klaser et al. 2008), structure features of interest points (Ryoo and Aggarwal 2009), trajectories (Wang et al. 2013; Raptis and Soatto 2010; Ni et al. 2015), holistic body shape (Liu et al. 2008), or key poses (Raptis and Sigal 2013), etc. Temporal evolution of human actions was captured in Fernando et al. (2015) and Kong et al. (2014). Bag-of-words (BoW) model is a common method for encoding these features in a vector format. BoW model first detects local or global features [e.g., spatiotemporal interests (Dollar et al. 2005; Laptev 2005) or histograms of oriented gradient (Dalal and Triggs 2005)] from videos. Then a clustering method such as k-means is employed to quantize these features into so-called visual words. After that, a histogram of the words contained in a video is computed and is considered as the vector format representation for the video. All these approaches use hand-crafted features, which normally require expensive human labor and expert knowledge to design extraction algorithms.

Thanks to recent deep learning techniques, human actions can be effectively learned from low-level observations (Ji et al. 2013; Karpathy et al. 2014; Simonyan and Zisserman 2014; Vondrick et al. 2016; Ma et al. 2016; Fernando et al. 2016). Specifically, these approaches use convolutional neural networks (CNNs) to perform convolution operation on images and build representations for video frames in a vector format. However, they generally require a large number of training samples as the CNNs they use have thousands of parameters to be learned and are very complex. Another line of research attempts to describe human actions using high-level semantics (Liu et al. 2011; Kong et al. 2014), i.e., action attributes. These binary action attributes explain whether a

particular motion pattern is observed in a video, such as “arm raise up” and “leg move forward”.

Human interaction recognition (Kong et al. 2014; Lan et al. 2012; Ryoo and Aggarwal 2009; Marszałek et al. 2009) and human-object interaction (Zhou et al. 2015) were also explored in recent years. Previous studies Ryoo and Aggarwal (2009) and Marszałek et al. (2009) recognize interactions in the same way as single-person action recognition approaches (Dollar et al. 2005; Laptev 2005). Specifically, an interactions was represented as a motion descriptor including all the people in a video, and then an action classifier was adopted to classify this interaction. Context information was exploited in Kong et al. (2014) and Lan et al. (2012) in order to capture the motion relationships between people. The context information between a pair of motion attributes was captured in Kong et al. (2014). They described human interactions by this context information, which was called interactive phrases in their work. Action context between individuals was modeled in Lan et al. (2012). Their model can automatically determine which two individuals are having interactions.

2.3 Feature Learning

Feature learning methods (Kobayashi 2014; Pirsiavash et al. 2009; Argyriou et al. 2008; Xu et al. 2014) have been proposed to learn better feature representations for the recognition task. The methods in Pirsiavash et al. (2009) and Kobayashi (2014) adopt linear projections to learn better features in a matrix form for classification. They reduce the degree of freedom of the model parameter matrix by decomposing it into two parts and enforcing a hard restriction on their rank. Different from them, we elegantly use features from two modalities for recognition. In addition, we use an effective information measure to produce more compact cross-modal features. The work in Argyriou et al. (2008) learned a few common features across tasks using a regularizer, which couples the tasks and enforces sparsity.

Multimodal approaches (Jia et al. 2014; Xie and Xing 2013; Xu et al. 2014) attempt to discover common features between features of various modalities. The work in Jia et al. (2014) treated RGB and depth videos as two modalities. They used a cross-modality regularizer to link the two modalities in order to deal with the missing modality problem. A multimodal metric learning method in Xie and Xing (2013) embeds data of arbitrary modalities into a single latent space. The optimal distance metric is learned to better measure the similarity between data of different modalities. The method in Xu et al. (2014) extends information bottleneck (Tishby et al. 1999) to a multi-view model. Multiple information sources are filtered through a “bottleneck”, and then a margin maximization approach is used to strengthen the discrimination of the model.

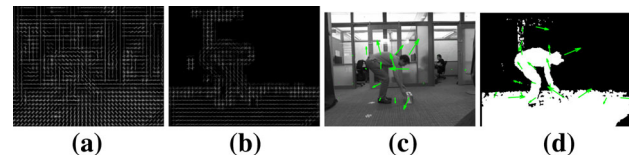


Fig. 3 Examples of HOG feature computed on **a** a color image and **b** a depth image, and optical flow computed on **c** a color image and **d** a depth image

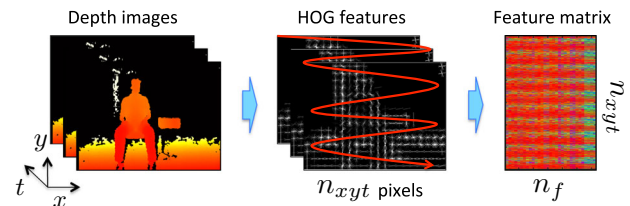


Fig. 4 Feature matrix of size $n_{xyt} \times n_f$ is constructed from features (e.g., HOG) computed on all the frames. n_{xyt} is the total number of pixels in all the feature frames, and n_f is the dimensionality of each local feature

Deep models have received lots of attentions in recent years, and various deep models have been developed for multi-modal learning (Ngiam et al. 2011; Andrew et al. 2013; Srivastava and Salakhutdinov 2014; Wang et al. 2015). The SplitAE method in Ngiam et al. (2011) assumes that a shared representation can be extracted from a single view, and can be used to reconstruct all views. Deep canonical correlation analysis (DCCA) was proposed in Andrew et al. (2013) to learn the correlations between two views using a deep architecture of nonlinear transformations. A multimodal deep Boltzmann machine (DBN) was presented in Srivastava and Salakhutdinov (2014). Their model uses modality-specific DBNs to build a layer of joint representation to fuse features from two modalities. The work in Wang et al. (2015) combined SplitAE and DCCA, and proposed three extensions for multimodal learning.

3 Max-Margin Heterogeneous Information Machine

The goal of this work is to utilize heterogeneous features from RGB-D action videos, and learn compact yet discriminative features for action recognition. Denote N RGB-D action videos for training purpose by $\{X_i, y_i\}_{i=1}^N$, where $X_i = \{X_i^{[v]}, X_i^{[z]}\} \in \mathcal{X}$ contains a RGB visual feature matrix $X_i^{[v]} \in \mathcal{X}_v$ and a depth feature matrix $X_i^{[z]} \in \mathcal{X}_z$ extracted from RGB-D data, and $y_i \in \mathcal{Y}$ is the corresponding action label. Each feature matrix contains both HOG and HOF feature descriptors (see Fig. 3). Note that $X_i^{[v]}$ and $X_i^{[z]}$ in our work are defined as feature matrices of size $n_{xyt} \times n_f$ (see Fig. 4), different from feature vectors (such as bag-of-word

models) containing $n_{xyt} \times n_f$ elements that are popularly used in the computer vision community. In this work, features $X_i^{[v]}$ and $X_i^{[z]}$ are extracted from a spatiotemporal grid of $n_{xyt} = n_x \times n_y \times n_t$, and n_f is the dimensionality of each local feature. HOG and HOF features are concatenated horizontally, i.e., n_f is the dimensionality of concatenated features. Action representation in a matrix form allows us to capture inherent structure of features, such as spatiotemporal relationships. However, these relationships are collapsed in a vector form feature representation. Note that one can pull out other dimensions rather than the feature dimension in $X_i^{[v]}$ and $X_i^{[z]}$, but the structure of n_{xyt} pixels in the feature matrices will not be conserved by the proposed model.

RGB-D action data X_i contain two modalities, visual features $X_i^{[v]}$ and depth features $X_i^{[z]}$. The major challenge for effectively using the two-modality features is that they come from different distributions, and thus their similarities could not be measured directly. To solve this problem, we would like to learn two projection functions $W_o^{[v]}$ and $W_o^{[z]}$ for visual features $X_i^{[v]}$ and depth features $X_i^{[z]}$, respectively. Each of the projection functions maps the corresponding features to a space \mathcal{O} shared between the two modalities: $W_o^{[v]} : \mathcal{X}_v \rightarrow \mathcal{O}$, and $W_o^{[z]} : \mathcal{X}_z \rightarrow \mathcal{O}$. After learning the projection functions, a classification model W_w can be learned to classify given the learned shared features $O \in \mathcal{O}$.

The learned shared features O may not capture all the discriminative information for classification. Some important cues are not shared between modalities. We takes this into account by introducing private features for each modality. Two modality-specific projection matrices $W_q^{[v]}$ and $W_q^{[z]}$ are adopted to learn private features $Q^{[v]} \in \mathcal{Q}_v$ and $Q^{[z]} \in \mathcal{Q}_z$

from the original visual and depths features, respectively: $W_q^{[v]} : \mathcal{X}_v \rightarrow \mathcal{Q}_v$, and $W_q^{[z]} : \mathcal{X}_z \rightarrow \mathcal{Q}_z$. Classification models $W_w^{[v]}$ and $W_w^{[z]}$ can also be learned given the learned private features $Q^{[v]}$ and $Q^{[z]}$.

In this work, we learn all the projection matrices and classification models simultaneously. Therefore, the learned projections are optimized for classification. We focus on learning a discriminant function $F : \mathcal{X} \times \mathcal{Y} \rightarrow \mathcal{R}$ that scores each training sample (X_i, y_i) . The function F is applied to compute the compatibility among the input RGB-D features X_i , the learned cross-modal features O , the private features Q , and the action label y_i . A list of mathematical symbols used this paper is given in Table 1.

3.1 Model Formulation

Suppose we are given M types of modalities $X_i^{[m]}|_{m=1}^M$. Here, m is the index of modality, which can be either visual ($m = 1$ or $m = v$) or depth ($m = 2$ or $m = z$). We represent both of the two modality features in a matrix form in order to capture inherent spatiotemporal structure. In this paper, we are interested in a binary linear discriminant function $F(X_i, y|W) = \text{Tr}(W^T X_i) = \sum_{m=1}^M \text{Tr}(W^{[m]T} X_i^{[m]})$ parameterized by a model weight matrix W . In this work, we learn both shared features and private features from visual and depth modalities in order to capture rich discriminative information for classification. The graphical illustration of our model is shown in Figure 5. The one-vs-one scheme is adopted to extend our binary classifiers to a multi-class classifier.

Table 1 List of mathematical symbols

Variable	Size	Meaning
m	Scalar	Modality indicator. m can be visual ($m = v$ or $m = 1$) or depth modality ($m = z$ or $m = 2$).
n_{xyt}	Scalar	Number of spatiotemporal feature points.
n_f	Scalar	Dimensionality of each feature point.
d_o	Scalar	Dimensionality of the shared features O .
d_q	Scalar	Dimensionality of the private features $Q^{[m]}$.
$X^{[m]}$	$n_{xyt} \times n_f$	Feature matrix of modality m .
O	$n_{xyt} \times d_o$	The learned shared features.
$Q^{[m]}$	$n_{xyt} \times d_q$	The learned private features of modality m .
$W_o^{[m]}$	$n_f \times d_o$	Projection matrix of modality m for learning shared features.
$W_q^{[m]}$	$n_f \times d_q$	Projection matrix of modality m for learning private features.
W_w	$n_{xyt} \times d_o$	Classification matrix for the shared features O .
$W_w^{[m]}$	$n_{xyt} \times d_q$	Classification matrix for the private features $Q^{[m]}$ of modality m .
$W_O^{[m]}$	$n_{xyt} \times d_f$	Parameter matrix for extracting shared features and classifying the features. $W_O^{[m]} = W_w W_o^{[m]}$.
$W_Q^{[m]}$	$n_{xyt} \times d_f$	Parameter matrix for extracting private features and classifying the features. $W_Q^{[m]} = W_w^{[m]} W_q^{[m]}$.

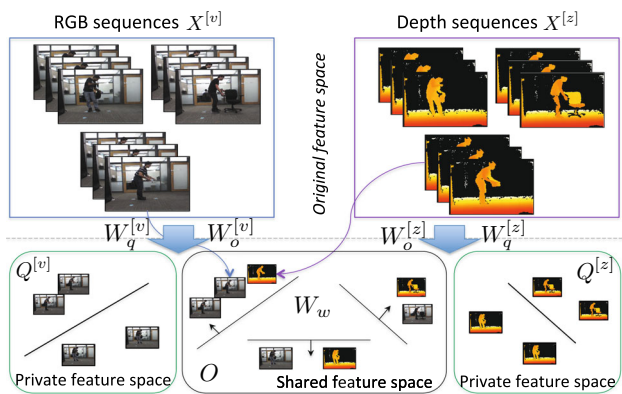


Fig. 5 Graphical illustration of the proposed MMHIM model. Parameter matrix $W_o^{[m]}$ ($m = 1, \dots, M$) projects the m modality data, $X^{[m]}$, into a learned shared space, and $W_q^{[m]}$ ($m = 1, \dots, M$) projects the data $X^{[m]}$ into private spaces. Classification is performed using both the learned shared features and private features

Shared information between visual and depth modalities captures complex correlations between them. Model parameter $W_o^{[m]}$ is used to extract shared information and classify actions. One of the challenges in RGB-D action recognition is that the two modalities, visual features and depth features, are in different feature spaces, and thus their similarities cannot be directly computed. We solve this problem by decomposing the model parameter $W_o^{[m]}$ into two components, W_w and $W_o^{[m]}$: $W_o^{[m]} = W_w W_o^{[m]T}$, which induces a bilinear model (Pirsiavash et al. 2009). Parameter matrix $W_o^{[m]} \in \mathcal{R}^{n_f \times d_o}$ ($m = 1, \dots, M$) projects the m -th modality data, $X^{[m]}$, onto a learned shared space, and parameter matrix $W_w \in \mathcal{R}^{n_{xyt} \times d_o}$ is applied to classify the projected data regardless of the modality. W_w is a spatiotemporal template defined over d_o features at each spatiotemporal location. Obviously, the rank of the model parameter matrix $W_o^{[m]}$ will be enforced to be at most d_o .

In addition to the shared features, each modality may also contain discriminative information but cannot be shared with the other modality. We capture such private features of modality m for classification using model parameter $W_q^{[m]}$. Similar to $W_o^{[m]}$ for shared features, $W_q^{[m]}$ is also decomposed into two components, $W_w^{[m]}$ and $W_q^{[m]}$. Parameter matrix $W_q^{[m]} \in \mathcal{R}^{n_f \times d_q}$ projects the original low-level data of modality m to a low-dimensional space, and $W_w^{[m]} \in \mathcal{R}^{n_{xyt} \times d_q}$ is used to classify the projected data, which is essentially a spatiotemporal template for the projected data of d_q features. The rank of the model parameter $W_q^{[m]}$ is enforced to be at most d_q .

Once the optimal model parameter matrix W is learned from training data, the action label y_i^* of a sample X_i can be computed by

$$\begin{aligned}
 y_i^* &= \text{sign} \left[\text{Tr}(W^T X_i) \right] = \text{sign} \left[\text{Tr}(W_o^T X_i + W_q^T X_i) \right] \\
 &= \text{sign} \left[\sum_m \text{Tr} \left(\underbrace{W_o^{[m]} W_w^{[m]T} X_i^{[m]}}_{\text{shared information}} \right. \right. \\
 &\quad \left. \left. + \underbrace{W_q^{[m]} W_w^{[m]T} X_i^{[m]}}_{\text{private information}} \right) \right], \tag{1}
 \end{aligned}$$

where $\text{sign}(\cdot)$ is the sign function.

We train the MMHIM in Eq. (1) in a max-margin framework. Based on the empirical risk minimization principle, we formulate our learning problem as

$$\begin{aligned}
 \min_{W_w, W_w^{[m]}, W_o^{[m]}, W_q^{[m]}} & \lambda \cdot r(W_w, W_w^{[m]}, W_o^{[m]}, W_q^{[m]}) \Big|_m \\
 & + \phi(W_o^{[m]}, W_q^{[m]}) \Big|_m + \eta \cdot l(W_w, W_w^{[m]}, W_o^{[m]}, W_q^{[m]}) \Big|_m. \tag{2}
 \end{aligned}$$

For succinctness, “ $|_m$ ” indicates that the parameters of the two modalities ($m = v$ and $m = z$) are jointly optimized. $\phi(\cdot)$ is a regularizer for reducing noise in the projected data, $r(\cdot)$ is an additional regularization term related to the margin of our bilinear model, and $l(\cdot)$ computes training loss for the two-modality data. λ and η are trade-off parameters balancing the importance of the corresponding terms.

3.1.1 Margin Regularizer $r(W_w, W_w^{[m]}, W_o^{[m]}, W_q^{[m]})|_m$

Regularizer $r(W_w, W_w^{[m]}, W_o^{[m]}, W_q^{[m]})|_{m \in \{v, z\}}$ is used to measure the margin of the bilinear classifier. Minimizing $r(W_w, W_w^{[m]}, W_o^{[m]}, W_q^{[m]})|_{m \in \{v, z\}}$ is equivalent to maximizing the margin of the bilinear model, thereby improving the generalization power. The margin regularizer $r(\cdot)$ is defined as

$$\begin{aligned}
 r(W_w, W_w^{[m]}, W_o^{[m]}, W_q^{[m]}) \Big|_m & x \\
 &= \frac{1}{2} \left[\underbrace{r_o(W_w, W_w^{[m]}, W_o^{[m]}, W_q^{[m]}) \Big|_m}_{\text{margin for shared features}} \right. \\
 &\quad \left. + \underbrace{r_q(W_w, W_w^{[m]}, W_o^{[m]}, W_q^{[m]}) \Big|_m}_{\text{margin for private features}} \right], \tag{3}
 \end{aligned}$$

$$\begin{aligned}
 r_o(W_w, W_w^{[m]}, W_o^{[m]}, W_q^{[m]}) \Big|_m & \\
 &= \frac{1}{2} \left[\text{Tr}(W_w W_o^{[v]T} W_o^{[v]} W_w^T) \right. \\
 &\quad \left. + \text{Tr}(W_w W_o^{[z]T} W_o^{[z]} W_w^T) \right], \tag{4}
 \end{aligned}$$

$$\begin{aligned}
 r_q(W_w, W_w^{[m]}, W_o^{[m]}, W_q^{[m]}) \Big|_m &= \frac{1}{2} \\
 &\times \left[\text{Tr}(W_w^{[m]} W_q^{[v]T} W_q^{[v]} W_w^{[m]T}) \right.
 \end{aligned}$$

$$+ \text{Tr}(W_w^{[m]} W_q^{[z]T} W_q^{[z]} W_w^{[m]T})]. \tag{5}$$

The regularization term $r(W_w, W_w^{[m]}, W_o^{[m]}, W_q^{[m]})|_m$ considers the classifier margins of the models for both shared features and private features, $r_o(\cdot)$ and $r_q(\cdot)$. It naturally induces low-rank classifiers with the maximum rank of d_o and d_q for the two types of features, respectively. This restricts the degree of freedom of model parameter matrices. As shown in Wolf et al. (2007), the VC-dimension of low-rank classification models was proven to be less than that of the concatenated linear models.

Regularizer $r(W_w, W_w^{[m]}, W_o^{[m]}, W_q^{[m]})|_m$ is minimized to extract discriminative information from both cross-modal features O and private features Q for action recognition. It works together with $\phi(W_o^{[m]}, W_q^{[m]})|_m$ in Eq. (6) to extract discriminative information and filter out noise for the recognition task.

3.1.2 Projection Regularizer $\phi(W_o^{[m]}, W_q^{[m]})|_m$

Regularizer $\phi(W_o^{[m]}, W_q^{[m]})|_m$ is a function that attempts to summarize and compress the original two-modality data. Since the raw RGB and depth data may not be in the same space, we use this term to compress the data, and discover shared and modality-specific knowledge between the two modalities. We define this term as

$$\begin{aligned} \phi(W_o^{[m]}, W_q^{[m]})|_m = & \left[\underbrace{\phi_o(W_o^{[m]}, W_q^{[m]})|_m}_{\text{shared information}} \right. \\ & + \underbrace{\phi_q(W_o^{[m]}, W_q^{[m]})|_m}_{\text{private information}} \\ & \left. + \underbrace{\phi_r(W_o^{[m]}, W_q^{[m]})|_m}_{\text{redundant information}} \right], \tag{6} \end{aligned}$$

$$\phi_o(W_o^{[m]}, W_q^{[m]})|_m = I(X^{[v]}, O) + I(X^{[z]}, O), \tag{7}$$

$$\phi_q(W_o^{[m]}, W_q^{[m]})|_m = I(X^{[v]}, Q^{[v]}) + I(X^{[z]}, Q^{[z]}), \tag{8}$$

$$\phi_r(W_o^{[m]}, W_q^{[m]})|_m = I(Q^{[v]}, O) + I(Q^{[z]}, O), \tag{9}$$

where $I(\cdot, \cdot)$ computes mutual information between two variables J and K :

$$I(J, K) = \sum_j \sum_k p(j)p(k|j) \log \frac{p(k|j)}{p(k)}. \tag{10}$$

$X^{[m]} = \{X_i^{[m]}\}_{i=1}^N$ ($m = v$ or $m = z$) represents a set of all training samples in the m -th modality, $O = \frac{1}{2}(X^{[v]}W_o^{[v]} + X^{[z]}W_o^{[z]}) \in \mathcal{O}$ is the learned low-dimensional cross-modal features in the shared space, $Q^{[m]} = X^{[m]}W_q^{[m]}$ denotes the private features of the m -th modality.

Equations (7) and (8) are utilized to introduce cross-modality and private knowledge to the model through the learning of the shared features O and the private features Q . Cross-modal features O inherit information from both RGB and depth data, including motion, structure, and spatiotemporal relationship information. Private features, on the other hand, capture information that is not sharable, such as distance cue in the depth data. We show in the experiments that the learned features play an important role in the recognition of RGB-D actions and in case of missing one modality in training or testing phase. Equation (9) aims at reduce redundancies between the shared and private features.

In addition, the term $\phi(W_o^{[m]}, W_q^{[m]})|_m$ helps to reduce noise and produce compact representations for cross-modal features O and private features Q . In the learning of the features O and Q , a large amount of noise irrelevant to action labels would also be introduced to low-dimensional spaces, and thus degrades the recognition performance. By minimizing $\phi(W_o^{[m]}, W_q^{[m]})|_m$, both noisy and discriminative information in O and Q will be reduced, but the later one can be well captured by the regularizer $r(W_w, W_w^{[m]}, W_o^{[m]}, W_q^{[m]})|_m$ in Eq. (3). Parameter λ for regularizer $r(W_w, W_w^{[m]}, W_o^{[m]}, W_q^{[m]})|_m$ is used for balancing the importance of the noise filter in MMHIM.

3.1.3 Loss Function $l(W_w, W_w^{[m]}, W_o^{[m]}, W_q^{[m]})|_m$

Loss function $l(W_w, W_w^{[m]}, W_o^{[m]}, W_q^{[m]})|_m$ computes training loss given the learned model parameter matrices. We consider binary classifiers in this work, and define a hinge loss function $h(y, f(x)) = \max(0, 1 - yf(x))$ for each modality, which is similar to the one in the binary SVM:

$$l(W_w, W_w^{[m]}, W_o^{[m]}, W_q^{[m]})|_m \tag{11}$$

$$= \sum_i \left[h(y_i, \text{Tr}(W_o^{[v]} W_w^T X_i^{[v]})) + h(y_i, \text{Tr}(W_o^{[z]} W_w^T X_i^{[z]})) \right] \tag{12}$$

$$+ h(y_i, \text{Tr}(W_q^{[v]} W_w^{[v]T} X_i^{[v]})) + h(y_i, \text{Tr}(W_q^{[z]} W_w^{[z]T} X_i^{[z]}))]. \tag{13}$$

Here, the losses in Eq. (12) are incurred by the shared features O , and the ones in Eq. (13) are incurred by the private features Q .

3.1.4 Learning Formulation

Plugging Eqs. (3), (6), and (11) into Eq. (2), optimal parameter matrices $\{W_o^{[m]}, W_q^{[m]}, W_w^{[m]}, W_w\}|_m$ can be learned by the following constrained optimization problem:

$$\begin{aligned}
& \min_{W_w, W_w^{[m]}, W_o^{[m]}, W_q^{[m]}} \sum_m \left\{ I(X^{[m]}, O) + I(X^{[m]}, Q^{[m]}) \right. \\
& + I(Q^{[m]}, O) + \frac{1}{4} \lambda \left[\text{Tr}(W_w W_o^{[m]T} W_o^{[m]} W_w^T) \right. \\
& \left. \left. + \text{Tr}(W_w^{[m]} W_q^{[m]T} W_q^{[m]} W_w^{[m]T}) \right] + \eta \cdot \sum_i (\xi_i^{[m]} + \epsilon_i^{[m]}) \right\}, \\
\text{s.t. } & y_i \text{Tr}(W_o^{[m]} W_w^T X_i^{[m]}) \geq 1 - \xi_i^{[m]}, \quad \forall i, \forall m, \\
& y_i \text{Tr}(W_q^{[m]} W_w^{[m]T} X_i^{[m]}) \geq 1 - \epsilon_i^{[m]}, \quad \forall i, \forall m, \\
& \xi_i^{[m]} \geq 0, \quad \forall i, \forall m, \\
& \epsilon_i^{[m]} \geq 0, \quad \forall i, \forall m,
\end{aligned} \tag{14}$$

where $\xi_i^{[m]}$ and $\epsilon_i^{[m]}$ are slack variables for the shared features and the private features of the m -th modality in the i -th RGB-D video, respectively.

3.2 Model Learning

The above constrained optimization problem can be solved by a coordinate descent algorithm that solves for one set of parameter matrices at each step with the others fixed. Each step in the algorithm is a regularized risk minimization problem, which can be solved using a bundle method¹ (Teo et al. 2007; Do and Artieres 2009). In a nutshell, the bundle algorithm iteratively builds an increasingly accurate piecewise quadratic lower bound of the objective function. We adopt the bundle method as the inner problem solver due to its efficiency and good convergence.

We first reformulate the optimization problem (14) as an unconstrained regularized risk minimization problem:

$$\min_{W_w, W_w^{[m]}, W_o^{[m]}, W_q^{[m]}} \eta \cdot \sum_i \sum_m L_i^{[m]} + \sum_m R^{[m]}, \tag{15}$$

where

$$\begin{aligned}
L_i^{[m]} &= \max \left(0, 1 - y_i \text{Tr}(W_o^{[m]} W_w^T X_i^{[m]}) \right) \\
&+ \max \left(0, 1 - y_i \text{Tr}(W_q^{[m]} W_w^{[m]T} X_i^{[m]}) \right) \\
R^{[m]} &= I(X^{[m]}, O) + I(X^{[m]}, Q^{[m]}) + I(Q^{[m]}, O) + \frac{1}{2} \lambda \cdot \\
&\left[\text{Tr}(W_w W_o^{[m]T} W_o^{[m]} W_w^T) \right. \\
&\left. + \text{Tr}(W_w^{[m]} W_q^{[m]T} W_q^{[m]} W_w^{[m]T}) \right],
\end{aligned} \tag{16}$$

are empirical loss and regularizers, respectively. We solve this optimization problem using a coordinate descent algorithm that iteratively update one variable at a time.

¹ <https://forge.lip6.fr/projects/nrbm>

Update W_w . Specifically, if $\{W_w^{[m]}, W_o^{[m]}, W_q^{[m]}\}$ are fixed, the optimization problem is

$$\begin{aligned}
& \min_{W_w} \frac{1}{2} \lambda \sum_m \text{Tr}(W_w W_o^{[m]T} W_o^{[m]} W_w^T) \\
& + \eta \sum_i \sum_m \max \left(0, 1 - y_i \text{Tr}(W_o^{[m]} W_w^T X_i^{[m]}) \right).
\end{aligned} \tag{17}$$

To efficiently solve this problem, we define $A = \sum_m W_o^{[m]T} W_o^{[m]}$, and define two auxiliary variables $\widehat{W}_w = W_w A^{\frac{1}{2}}$ and $\widehat{X}_i^{[m]} = X_i W_o^{[m]} A^{-\frac{1}{2}}$. Note that A is a matrix of size $d \times d$ that is in general invertible for small d . Then the problem (17) can be equivalently rewritten as

$$\min_{\widehat{W}_w} \frac{1}{2} \lambda \text{Tr}(\widehat{W}_w^T \widehat{W}_w) + \eta \sum_i \sum_m \max \left(0, 1 - y_i \text{Tr}(\widehat{W}_w^T \widehat{X}_i^{[m]}) \right). \tag{18}$$

This is an unconstrained regularized risk minimization problem equivalent to linear SVM if \widehat{W}_w and $\widehat{X}_i^{[m]}$ are vectorized. We solve this problem using a bundle method. After learning \widehat{W}_w , the original parameter matrix W_w can be reconstructed by $W_w = \widehat{W}_w A^{-\frac{1}{2}}$.

Update $W_w^{[m]}$. We fix $\{W_w, W_o^{[m]}, W_q^{[m]}\}$, and solve

$$\begin{aligned}
& \min_{W_w^{[m]}} \frac{1}{2} \lambda \sum_m \text{Tr}(W_w^{[m]} W_q^{[m]T} W_q^{[m]} W_w^{[m]T}) \\
& + \eta \sum_i \sum_m \max \left(0, 1 - y_i \text{Tr}(W_q^{[m]} W_w^{[m]T} X_i^{[m]}) \right).
\end{aligned} \tag{19}$$

Similar to the optimization procedure of parameter matrix W_w , we also define $B = W_q^{[m]T} W_q^{[m]}$, and introduce two auxiliary variables $\overline{W}_w^{[m]} = W_w^{[m]} B^{\frac{1}{2}}$ and $\overline{X}_i^{[m]} = X_i^{[m]} W_q^{[m]} B^{-\frac{1}{2}}$. Then the optimization problem (19) can be equivalently given by

$$\begin{aligned}
& \min_{\overline{W}_w^{[m]}} \frac{1}{2} \lambda \sum_m \text{Tr} \left(\overline{W}_w^{[m]T} \overline{W}_w^{[m]} \right) \\
& + \eta \sum_i \sum_m \max \left(0, 1 - y_i \text{Tr}(\overline{W}_w^{[m]T} \overline{X}_i^{[m]}) \right).
\end{aligned} \tag{20}$$

This is also an unconstrained regularized risk minimization problem. If $\overline{W}_w^{[m]}$ and $\overline{X}_i^{[m]}$ are vectorized, this problem can be solved using standard linear SVM solver. After learning $\overline{W}_w^{[m]}$, the original parameter matrix $W_w^{[m]}$ can be reconstructed by $W_w^{[m]} = \overline{W}_w^{[m]} B^{-\frac{1}{2}}$.

Update $W_o^{[m]}$. When $\{W_w, W_w^{[m]}, W_q^{[m]}\}$ are fixed, $W_o^{[m]}$ for each modality can be optimized in a similar form to Eq. (15)

and (16) but with W_w as constant. We define $C = W_w^T W_w$, and further define two auxiliary variables, \tilde{W}_o and \tilde{X}_i , as $\tilde{W}_o^{[m]} = W_o^{[m]} C^{\frac{1}{2}}$ and $\tilde{X}_i^{[m]} = X_i^{[m]T} W_w C^{-\frac{1}{2}}$. Then, the parameter matrix $\tilde{W}_o^{[m]}$ for each modality can be optimized independently by

$$\min_{\tilde{W}_o^{[m]}} \frac{1}{2} \text{Tr} \left(\tilde{W}_o^{[m]T} \tilde{W}_o^{[m]} \right) + \lambda I(\tilde{X}^{[m]}, \tilde{O}) + \eta \sum_i \max \left(0, 1 - y_i \text{Tr}(\tilde{W}_o^{[m]T} \tilde{X}_i^{[m]}) \right), \tag{21}$$

with the assumption that the conditional distribution $p(W_w, C^{-\frac{1}{2}} | X^{[m]}, O)$ is a uniform distribution². This is also an unconstrained regularized risk minimization problem and can be solved by a bundle algorithm if $\tilde{W}_o^{[m]}$ and $\tilde{X}_i^{[m]}$ are unfolded into vectors. We repeat this step twice, each of which is fed with visual features $X_i^{[v]}$ or depth feature $X_i^{[z]}$. After optimizing $\tilde{W}_o^{[m]}$, $W_o^{[m]}$ can be recovered by $W_o^{[m]} = \tilde{W}_o^{[m]} C^{-\frac{1}{2}}$.

Update $W_q^{[m]}$. When $\{W_w, W_w^{[m]}, W_o^{[m]}\}$ are fixed, $W_q^{[m]}$ can be optimized by

$$\min_{W_q^{[m]}} \frac{1}{2} \text{Tr} \left(W_w^{[m]} W_q^{[m]T} W_q^{[m]} W_w^{[m]T} \right) + \lambda I(X^{[m]}, Q^{[m]}) + \lambda I(Q^{[m]}, O) + \eta \sum_i \max \left(0, 1 - y_i \text{Tr}(W_q^{[m]} W_w^{[m]T} X_i^{[m]}) \right), \tag{22}$$

We use the similar method that is used in learning $W_o^{[m]}$. We define $D = W_w^{[m]T} W_w^{[m]}$, and further introduce two auxiliary variables $\hat{W}_q^{[m]} = W_q^{[m]} D^{\frac{1}{2}}$ and $\hat{X}_i^{[m]} = X_i^{[m]T} W_w^{[m]} D^{-\frac{1}{2}}$. Then the parameter matrix $W_q^{[m]}$ for modality m can be learned by

$$\min_{\hat{W}_q^{[m]}} \frac{1}{2} \text{Tr} \left(\hat{W}_q^{[m]T} \hat{W}_q^{[m]} \right) + \lambda I(\hat{X}^{[m]}, \hat{Q}^{[m]}) + \lambda I(\hat{Q}^{[m]}, \hat{O}) + \eta \sum_i \max \left(0, 1 - y_i \text{Tr}(\hat{W}_q^{[m]T} \hat{X}_i^{[m]}) \right), \tag{23}$$

The learning of parameter $\hat{W}_q^{[m]}$ can be solved using the bundle algorithm. One of the key steps in the bundle algorithm is computing the subgradient of the mutual information term $I(\hat{O}, \hat{Q}^{[m]})$ in the objective function in Eq. (23) with respect to the model parameter $\hat{W}_q^{[m]}$. In this work, the subgradient with respect to the (i, j) -th element in the model parameter $\hat{W}_q^{[m]}$ can be computed by

$$\frac{\partial I(\hat{O}, \hat{Q}^{[m]})}{\partial \hat{W}_{q(i,j)}^{[m]}} = \text{Tr} \left\{ \left[p(\hat{O}) \left(\log \frac{\hat{W}_o^{[m]+} \hat{W}_q^{[m]}}{p(\hat{Q}^{[m]})} + 1 \right) \right]^T \cdot \hat{W}_o^{[m]+} \cdot I_{ij} \right\}, \tag{24}$$

where $\hat{W}_o^{[m]+}$ computes the pseudo-inverse of $\hat{W}_o^{[m]}$: $\hat{W}_o^{[m]+} = \hat{W}_o^{[m]T} (\hat{W}_o^{[m]} \hat{W}_o^{[m]T})^{-1}$. I_{ij} is a matrix (of size $n_f \times d_p$) with all 0s but with 1 at (i, j) .

We solve the optimization problem (23) twice, each of which is fed with visual features $X_i^{[v]}$ or depth features $X_i^{[z]}$. After optimizing $\hat{W}_q^{[m]}$, $W_q^{[m]}$ can be recovered by $W_q^{[m]} = \hat{W}_q^{[m]} D^{-\frac{1}{2}}$.

The proposed MMHIM is solved by iteratively optimizing problems (18), (20), (21), and (23) until convergence. This is a biconvex problem as optimizing one parameter matrix holding the others fixed is a convex problem. The algorithm converges as optimizing each of model parameter matrices reduces objective function value.

3.3 Using Feature Vectors for Classification

The proposed approach takes feature matrices as input in order to capture spatiotemporal structures of human body parts. However, compared with features in vector format, there are not too many features represented in a matrix form. In order to utilize existing vector-based features [such as skeleton feature vectors (Du et al. 2015; Wang et al. 2012b) or normal vectors (Oreifej and Liu 2013; Yang et al. 2014)], we propose a generalized framework that utilizes features in both a vector format and a matrix format.

Skeleton features are popularly used in action recognition from RGB-D videos due to its high discriminative power (Du et al. 2015; Wang et al. 2012b; Xia and Aggarwal 2013). We extract skeleton feature using Du et al. (2015) in this work, represent it in a vector format, and use it for classification together with visual and depth cues that are defined in a matrix format. A new potential linear function $w_s^T x^{[s]}$ is added to the discriminant function in Eq. (1). Here, w_s is a vector of model parameters, and $x^{[s]}$ is a vector of skeleton features. The linear function $w_s^T x^{[s]}$ measures the compatibility between the skeleton features $x^{[s]}$ and the action label $+1/-1$. By adding $w_s^T x^{[s]}$, the new discriminant function is

$$y_i^* = \text{sign} \left[\text{Tr}(W_O^T X_i + W_Q^T X_i) + w_s^T x^{[s]} \right]. \tag{25}$$

Note that we do not learn a shared feature space for the skeleton feature in this work, different from visual and depth features where a shared feature space is learned.

The model parameter w_s can be jointly learned with other model parameters $\{W_o^{[m]}, W_q^{[m]}, W_w^{[m]}, W_w\}_m$ using the following optimization formulation similar to (15):

² Please refer to the supplemental material for details.

$$\min_{w_s, W_w, W_w^{[m]}, W_o^{[m]}, W_q^{[m]}} \eta \cdot \sum_i \sum_m L_i^{[m]} + \sum_m R^{[m]}, \quad (26)$$

where

$$\begin{aligned} L_i^{[m]} &= \max(0, 1 - y_i w_s^T x^{[s]}) \\ &+ \max\left(0, 1 - y_i \text{Tr}(W_o^{[m]} W_w^T X_i^{[m]})\right) \\ &+ \max\left(0, 1 - y_i \text{Tr}(W_q^{[m]} W_w^{[m]T} X_i^{[m]})\right) \\ R^{[m]} &= I(X^{[m]}, O) + I(X^{[m]}, Q^{[m]}) + I(Q^{[m]}, O) \\ &+ \frac{1}{2} \lambda \left[\|w\|^2 + \text{Tr}(W_w W_o^{[m]T} W_o^{[m]} W_w^T) \right. \\ &\left. + \text{Tr}(W_w^{[m]} W_q^{[m]T} W_q^{[m]} W_w^{[m]T}) \right]. \end{aligned} \quad (27)$$

The optimization problem can be solved by a coordinate descent algorithm similar to the one we proposed in Sect. 3.2. If we update w_s and fix all the other parameters, the learning problem can be written as

$$\min_{w_s} \frac{1}{2} \lambda \|w\|^2 + \max(0, 1 - y_i w_s^T x^{[s]}), \quad (28)$$

which is a standard linear SVM optimization problem and can be solved using an off-the-shelf SVM solver.

3.4 Structured Prediction Model

The main limitation of the above MMHIM model is that it cannot be used in structured prediction problems. In this work, we further extend the MMHIM to a structured prediction model that can capture the correlations between multiple outputs. We consider a special case of learning a multi-class MMHIM $\text{Tr}[W^T \Phi(X, y)]$ for n_c action categories. Here, W is model parameter matrix and $\Phi(X, y)$ is a feature function that models the agreement between low-level features X and action label y . Various structured prediction models can be developed based on Structured MMHIM by using more complex model structures (e.g., a sequence of video frames, multiple body parts in part-based models). This allows us to learn shared and private features, and structured labels jointly.

The key in the structured prediction model is the feature function $\Phi(X_i, y_i)$. In this work, we define the feature function as

$$W^T \Phi(X_i, y_i) = W_O^T \Phi_O(X_i, y_i) + W_Q^T \Phi_Q(X_i, y_i). \quad (29)$$

Similar to the ones in Eq. (1), both W_O^T and W_Q^T can be decomposed into two components, a classification component and a projection component. These components in W_O^T or W_Q^T are used to score the visual modality and depth modality in the sample X_i .

We define the potential functions $W_O^T \Phi_O(X, y)$ and $W_Q^T \Phi_Q(X, y)$ as

$$\begin{aligned} \text{Tr}[W_O^T \Phi_O(X, y)] &= \sum_m \sum_{a \in \mathcal{Y}} W_{w,a}^T \cdot 1(y = a) \cdot X^{[m]} W_o^{[m]}, \\ \text{Tr}[W_Q^T \Phi_Q(X, y)] &= \sum_m \sum_{a \in \mathcal{Y}} W_{w,a}^{[m]T} \cdot 1(y = a) \cdot X^{[m]} W_q^{[m]}, \end{aligned} \quad (30)$$

where a is an index for action labels and $1(\cdot)$ is an indicator function. Note that as we are considering a multi-class classification problem, the classification parameters W_w in W_O and $W_w^{[m]}$ in W_Q contain n_c classification templates, respectively: $W_w = [W_{w1}, W_{w2}, \dots, W_{wn_c}]$ and $W_w^{[m]} = [W_{w1}^{[m]}, W_{w2}^{[m]}, \dots, W_{wn_c}^{[m]}]$, where each W_{wt} or $W_{wt}^{[m]}$ can be regarded as a classifier for action category $t \in \{1, \dots, n_c\}$. Feature function $\Phi_O(X_i, y_i)$ can be explicitly expressed as

$$\Phi_O(X_i, y_i) = (\mathbf{0}, \mathbf{0}, \dots, X_i, \dots, \mathbf{0}, \mathbf{0})^T \in \mathcal{R}^{n_c \times d}. \quad (31)$$

Here, $\Phi_O(X_i, y_i)$ is a matrix with n_c rows. X_i locates in the y_i -th (out of n_c) row of $\Phi_O(X_i, y_i)$, and d is the length of X_i . Similarly, feature function $\Phi_Q(X_i, y_i)$ can be given by

$$\Phi_Q(X_i, y_i) = (\mathbf{0}, \mathbf{0}, \dots, X_i, \dots, \mathbf{0}, \mathbf{0})^T \in \mathcal{R}^{n_c \times d}. \quad (32)$$

Compared to MMHIM, Structured MMHIM only takes feature vectors as input and thus does not capture spatiotemporal structure information. Feature functions $\Phi_O(\cdot)$ and $\Phi_Q(\cdot)$ in Structured MMHIM are matrices corresponding to feature matrix X in MMHIM, and are projected using projection matrices $W_o^{[m]}$ and $W_q^{[m]}$, respectively. A straightforward way to design feature functions $\Phi_O(\cdot)$ and $\Phi_Q(\cdot)$ in Structured MMHIM is to follow the structured support vector machine (SSVM) (Joachims et al. 2009), where feature vectors are used in computing these feature functions. Following Joachims et al. (2009), feature vector X_i should be placed in the y_i -th row of $\Phi_O(X_i, y_i)$ (Eq. 31). Extensions to using feature matrix in Structured MMHIM is feasible but it requires the redesign of projection matrices and the classification parameter matrices.

We learn the Structured MMHIM using the formulation similar to the one proposed in (14), which is

$$\begin{aligned} \min_{W_w, W_w^{[m]}, W_o^{[m]}, W_q^{[m]}} \sum_m \left\{ I\left(\Phi(X^{[m]}, y), \Phi(O, y)\right) \right. \\ + I\left(\Phi(X^{[m]}, y), \Phi(Q^{[m]}, y)\right) + I\left(\Phi(Q^{[m]}, y), \Phi(O, y)\right) \\ + \frac{1}{2} \lambda \left[\text{Tr}\left(W_w W_o^{[m]T} W_o^{[m]} W_w^T\right) \right. \\ \left. + \text{Tr}\left(W_w^{[m]} W_q^{[m]T} W_q^{[m]} W_w^{[m]T}\right) \right] + \eta \sum_i (\xi_i^{[m]} + \epsilon_i^{[m]}) \left. \right\}, \end{aligned}$$

$$\begin{aligned}
 \text{s.t.} \quad & \text{Tr} \left(W_o^{[m]} W_w^T \Delta \Phi(X_i^{[m]}, y_i) \right) \geq 1 - \xi_i^{[m]}, \quad \forall i, \forall m, \\
 & \text{Tr} \left(W_q^{[m]} W_w^{[m]T} \Delta \Phi(X_i^{[m]}, y_i) \right) \geq 1 - \epsilon_i^{[m]}, \quad \forall i, \forall m, \\
 & \xi_i^{[m]} \geq 0, \quad \forall i, \forall m, \\
 & \epsilon_i^{[m]} \geq 0, \quad \forall i, \forall m,
 \end{aligned} \tag{33}$$

where $\Delta \Phi(X_i^{[m]}, y_i) = \Phi(X_i^{[m]}, y_i) - \Phi(X_i^{[m]}, y)$.

It should be noted that the key in Structured MMHIM is the feature functions $\Phi_O(\cdot)$ and $\Phi_Q(\cdot)$. Structured MMHIM is flexible in graph structure and is capable of predicting structured labels while MMHIM cannot. In the following, we design a Structured MMHIM for modeling temporal frames with varying length, which cannot be performed in MMHIM.

Example We use Structured MMHIM to model temporal frames in videos in this example. Consider a graph $G = \{V, E\}$, where V is a set of nodes and E is a set of edges. Here, a node corresponds to a (RGB or depth) video frame and an edge links two successive frames. Structured MMHIM projects the t -th RGB and depth video frames ($X_t^{[v]}$ and $X_t^{[z]}$) onto a shared feature space to learn shared features, and projects the two modalities onto independent private spaces to learn modality-specific features as well. Structured MMHIM predicts an action label $y_t \in \mathcal{Y}$ for each frame in a video using both shared and private features. A majority voting scheme is adopted on all the frames in a video to infer the label of the video.

Suppose a video has T frames, then the potential functions $W_o^T \Phi(X, y)$ and $W_q^T \Phi(X, y)$ in the Structured MMHIM can be defined as

$$\begin{aligned}
 & \text{Tr}[W_o^T \Phi_O(X, y)] \\
 &= \sum_m \left\{ \underbrace{\sum_{t=1}^T \sum_{a \in \mathcal{Y}} [W_{w,a}^T \cdot 1(y_t = a) \cdot X_t^{[m]} W_o^{[m]}]}_{\text{unary potential}} \right. \\
 & \left. + \underbrace{\sum_{t=1}^{T-1} \sum_{a \in \mathcal{Y}} \sum_{b \in \mathcal{Y}} [W_{w,a,b} \cdot 1(y_t = a) \cdot 1(y_{t+1} = b)]}_{\text{pairwise potential}} \right\},
 \end{aligned} \tag{34}$$

$$\begin{aligned}
 & \text{Tr}[W_q^T \Phi_Q(X, y)] \\
 &= \sum_m \left\{ \underbrace{\sum_{t=1}^T \sum_{a \in \mathcal{Y}} [W_{w,a}^{[m]T} \cdot 1(y = a) \cdot X_t^{[m]} W_q^{[m]}]}_{\text{unary potential}} \right. \\
 & \left. + \underbrace{\sum_{t=1}^{T-1} \sum_{a \in \mathcal{Y}} \sum_{b \in \mathcal{Y}} [W_{w,a,b}^{[m]} \cdot 1(y_t = a) \cdot 1(y_{t+1} = b)]}_{\text{pairwise potential}} \right\}.
 \end{aligned} \tag{35}$$

Here, $W_w = \{W_{w,a}\}_{a=1, \dots, n_c}$ and $W_w^{[m]} = \{W_{w,a}^{[m]}\}_{a=1, \dots, n_c}$ are classification matrices, $\{W_{w,a,b}\}_{a,b=1, \dots, n_c}$ and $\{W_{w,a,b}^{[m]}\}_{a,b=1, \dots, n_c}$ are classification parameters, and $W_o^{[m]}$ and $W_q^{[m]}$ are projection matrices for shared and private features, respectively.

The first terms in the above two equations are unary potential functions and the second terms are pairwise potentials. Unary potentials model the compatibility between the low-level projected frame features $X_t^{[m]} W_o^{[m]}$ (or $X_t^{[m]} W_q^{[m]}$) and the classification template $W_{w,a}^T$ (or $W_{w,a}^{[m]T}$); while the pairwise potentials capture the compatibility between successive frames. We refer to this model as **Structured MMHIM-2** and the original Structured MMHIM in Eq. (30) and Eq. (33) as **Structured MMHIM-1**.

Comparison Structured MMHIM-1 proposes a general framework for structured prediction. It is not defined on a graph structure. By extending its potential functions in Eq. (30) to Eq. (34) and Eq. (35), the new structured model, Structured MMHIM-2, defines a graph for modeling frame sequences. Eq. (34) projects all the depth and RGB frames onto a shared space, and learn modality-specific features at the same time. Eq. (35) models the correlations between successive frames in a video. Both Structured MMHIM-1 and MMHIM have to sample a fixed number of frames (10 frames in this work) in order to fix the size of input matrices; while Structured MMHIM-2 is not restricted to the number of frames in a video. Both Structured MMHIM-1 and MMHIM do not capture the correlations between successive frames in a video; while Structured MMHIM-2 considers the correlations. MMHIM takes feature matrices as input, while Structured MMHIM-1 and Structured MMHIM-2 take feature vectors as input.

3.5 Model Properties

We would like to discuss key properties of the proposed MMHIM here.

Matrix format feature representation used in this work naturally considers spatiotemporal motion structure. Recall that the feature matrix X is of size $n_{xyt} \times n_f$. It pulls apart the feature dimension from the collapsed spatiotemporal dimension (x - y - t). In such a representation, the spatiotemporal structure is kept by n_{xyt} pixels in the feature matrix X . Motion relationships of body parts also exist in the rows of X . After projection using W_o or W_q , the structure of n_{xyt} pixels in X and the motion relationships are still conserved in the projected feature matrix XW_o or XW_q , as W_o or W_q only operates on the columns of X .

However, if we use a feature vector \mathbf{x} instead of a feature matrix X , the spatiotemporal structure and the motion relationships between body parts will not be conserved in

the projected features $W_o^T \mathbf{x}$ or $W_q^T \mathbf{x}$. This is because all the elements in \mathbf{x} are involved in the projection. Even though the feature vector \mathbf{x} itself captures the structure information using a spatiotemporal pyramid, for example, the information will collapse after projection due to the involvement of other elements in \mathbf{x} .

The matrix form representation used in this work is different from the 4th-order tensor format in Pirsivash et al. (2009). Their method captures width, height, temporal extent and feature dimension of a spatiotemporal window. The rank restriction in their work forces a spatiotemporal template to be separable along the x , y , t axis. By comparison, our representation considers spatiotemporal structures jointly. We put spatiotemporal dimensions together, and pull out feature dimension in this work.

Low-rank bilinear model MMHIM naturally models feature matrices using two model parameter matrices W_o (or W_q) and W_w . The rank of the proposed model is minimized to provide a better generalization power (Wolf et al. 2007). We show in the experiments that such a bilinear model can learn complex mappings, and the performance is even better than deep models (Liu and Shao 2013).

Information measure This is computed in the process of data projection in order to compress data and reduce noise in the learned space. We validate its effectiveness in the experiments.

Cross-modal features Our MMHIM learns cross-modal features from RGB and depth data. The cross-modal features are discriminative for classification as they capture implicit correlations between RGB and depth data, and inherit the characteristics of them including motion, 3D structural, and spatiotemporal correlation information.

Knowledge transfer The learned projection matrix $W_o^{[m]}$ and $W_q^{[m]}$ transfers information from original data $X^{[m]}$ to the learned shared features O and private features Q . This helps exploit cross-modal knowledge if one modality is missing in testing.

Structured prediction Structured MMHIM is capable of predicting structured outputs. This allows us to fuse heterogeneous cues and capture relationships between multiple outputs at the same time. Structured MMHIMs can be used in temporal series domain while MMHIM cannot be.

The third modality Our MMHIM uses two modalities and can be extended to using the third modality. For Kinect sensors, the third modality could be the skeleton features, which capture motion information from body joints. However, existing skeleton features are generally represented in

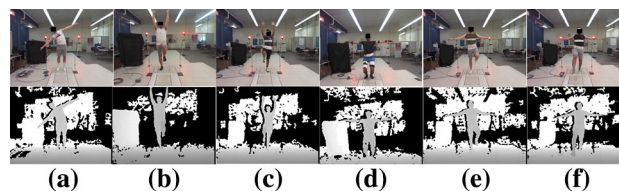


Fig. 6 Examples frames of NEU-UB dataset. The 6 action classes from left to right are **a** “Lateral Bend”, **b** “Left Leg Lift”, **c** “Right Leg Lift”, **d** “Squat”, **e** “Step Backward”, and **f** “Step Forward”

a vector format (Du et al. 2015; Wang et al. 2012b). In order to project the skeleton features into the feature space shared with visual and depth modalities, the raw skeleton features need to be represented in a matrix format so that it can be fed into MMHIM. A possible modification for changing vector-based skeleton features into matrix-based skeleton features can be made by: (1) Fixing the number of spatiotemporal locations of the features. (2) Pulling out the feature dimension. Consequently, the modified skeleton features are in matrix format and can be shared with the other two sources in the learned space. Nevertheless, this requires us to design a new feature representation for skeleton features, which is beyond the scope of this paper.

4 Experiments

4.1 Datasets and Settings

We collect a new RGB-D action dataset named NEU-UB dataset to test our approach. Example frames displayed in Fig. 6 show that depth videos in this dataset are extremely noisy. Therefore, it would be challenging for the methods that only use depth modality. This dataset was captured by three types of devices, including a Kinect sensor, Vicon sensors, and IMU sensors. We utilize the RGB and depth data generated by the Kinect sensor in this work. The NEU-UB dataset contains 6 action categories, including “Lateral Bend”, “Left Leg Lift”, “Right Leg Lift”, “Squat”, “Step Backward”, and “Step Forward”. Each action is performed by 20 subjects. Each actor repeats an action 5 times, to provide a total of 600 RGB-D videos. Visually similar temporal segments in different action categories frequently occur in this dataset, for example, “Step Backward” and “Step Forward”.

The proposed method is also evaluated on the MSR Action Pairs dataset (Oreifej and Liu 2013) and MSR Daily Activity dataset (Wang et al. 2012b). MSR Action Pairs dataset is an indoor RGB-D action dataset containing 12 types of activities performed by 10 subjects with both RGB and depth videos. Each actor repeats an action for three times, to provide a total of 360 videos for each of the RGB and depth modalities. MSR Daily Activity dataset contains 16 types of activities

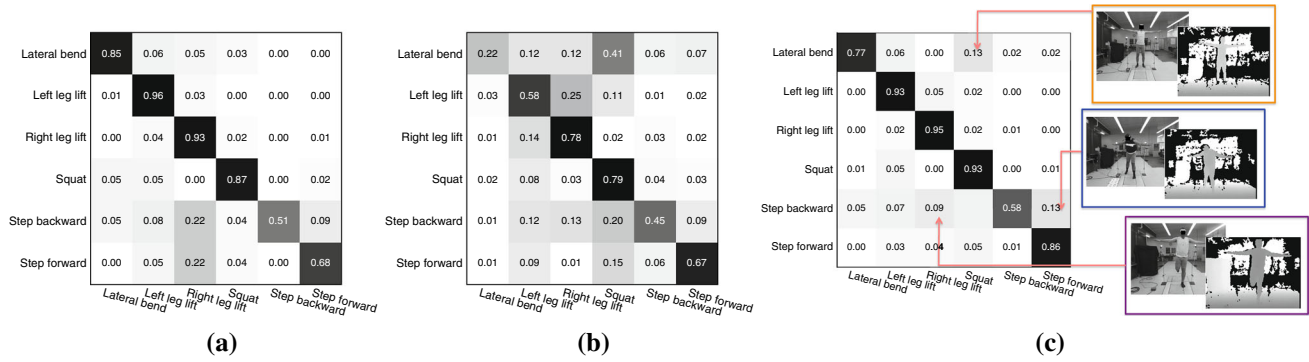


Fig. 7 Confusion matrix of **a** RGB-only (overall accuracy = 80.00%), **b** depth-only (overall accuracy = 58.25%), and **c** full MMHIM method (overall accuracy = 83.51%) on the NEU-UB dataset. Misclassification examples are also shown on the right

performed by 10 subjects. Each actor repeats an action twice, providing a total of 320 videos for each of the RGB and depth channels.

4.2 NEU-UB Action Dataset

Videos in this dataset are temporally normalized to 10 frames with spatial resolution of 120×160 . Histograms of oriented gradient feature and histograms of oriented flow feature are both extracted from color and depth videos in this dataset. A total of $n_{xyt} = 3000$ patches are extracted from each video, with the feature dimensionality of $n_f = 93$. The cross-validation training strategy is adopted for this dataset. The videos of the first 10 subjects are used for training, videos of 4 subjects are used for cross-validation, and the remaining videos of the other subjects are adopted for testing.

4.2.1 Classification Performance

Confusion matrices of the proposed MMHIM using RGB data, depth data, and the full RGB-D data on NEU-UB dataset are illustrated in Fig. 7. MMHIM achieves 83.51% accuracy in classifying actions in RGB-D videos. Misclassifications are mainly due to visually similar movements, for example, “Step backward” and “Step forward”, “Step backward” and “Right leg lift” (shown on the right hand side). 13% of “Lateral bend” videos are misclassified as “Squat”. This is mainly due to similar temporal segments in “Lateral bend”. The videos in “Lateral bend” has long durations with a person standing still, which is very similar to the ones in “Squat”. 13% of “Step backward” videos are misclassified as “Step forward” due to motion similarities. The two actions mainly differ in the distance changes from the human subject to the camera along temporal axis, which is not very clear in color videos. 7 and 9% of “Step backward” videos are also misclassified as “Left leg lift” and “Right leg lift”, respectively. The underlying reason is that in these misclassified videos,

Table 2 Recognition accuracy of comparison methods on NEU-UB dataset

Methods	Accuracy (%)
HON4D (Oreifej and Liu 2013)	58.42
SNV (Yang et al. 2014)	60.18
Linear SVM	75.44
LTTL (Jia et al. 2014)	77.89
Bilinear SVM (Pirsiavash et al. 2009)	80.70
BHIM (Kong and Fu 2015)	81.22
Structured MMHIM-1	82.63
MMHIM	83.51
Structured MMHIM-2	85.56
MMHIM+Skeleton	87.78

Bold value indicates the best result in comparison

human subjects perform other similar actions (lift his/her leg) during their action executions, which confuse MMHIM.

4.2.2 Comparison Results

We use Du et al. (2015) to extract skeleton features, and represent the features in a vector format. MMHIM+Skeleton method is built using the formulation in Sect.3.3. We compare MMHIM+Skeleton method, MMHIM, Structured MMHIM-1, Structured MMHIM-2, with existing methods (Pirsiavash et al. 2009; Jia et al. 2014; Oreifej and Liu 2013; Yang et al. 2014; Kong and Fu 2015) on NEU-UB dataset. Linear SVM is adopted as baseline. Comparison results are shown in Table 2.

Results show that MMHIM+Skeleton method outperforms all the comparison approaches. MMHIM+Skeleton method achieves 87.78% accuracy, outperforming MMHIM by 4.27% due to the use of skeleton features. Skeleton features naturally capture motion information of 20 joints, which is helpful in discriminating visually similar action categories. Structured MMHIM-2 outperforms MMHIM by 2.05% on NEU-UB dataset as it captures the correlation between two

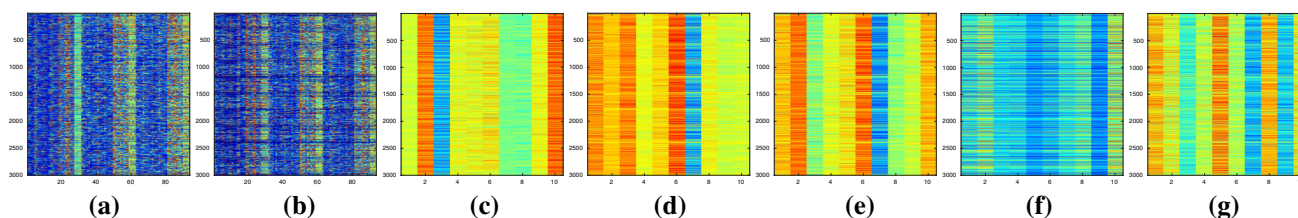


Fig. 8 Visualizations of **a** raw RGB features $X^{[v]}$, **b** raw depth features $X^{[z]}$, **c** the projected visual features $X^{[v]}W_o^{[v]}$, **d** the projected depth features $X^{[z]}W_q^{[z]}$, **e** the learned cross-modal features O in the shared space, **f** the projected visual features $X^{[v]}W_q^{[v]}$, and **g** the projected depth features $X^{[z]}W_o^{[z]}$

successive frames using pairwise potentials in Eq. (34) and Eq. (35). By contrast, MMHIM only samples 10 frames in a video and does not capture the correlation information. MMHIM achieves 83.51% accuracy, 2.29% higher than the BHIM with the same features. This demonstrates the benefit of learning private features in MMHIM. Given RGB-D data, MMHIM learns two extra pieces of information, i.e., private features for RGB data and private features for depth data. By encouraging minimum redundant information with the learned shared features, the learned private features capture unique information that is not sharable with the other modality. For example, this information can be a 2D edge cue in visual RGB modality and a distance cue in depth modality. Such information is discriminative for classification, and thus using this extra information can further improve the classification performance in MMHIM. MMHIM outperforms linear SVM due to the learning of better feature representations. Linear SVM simply concatenates visual and depth features into a long vector, which may not be able to capture complex correlations between visual and depth modalities. However, MMHIM finds the optimal space for fusing the two modalities, and thus improves the performance. Similar to MMHIM, bilinear SVM also learns a shared feature space for the two modalities. Nevertheless, MMHIM filters out redundant information and reduces noise in the learned features using the information measure $\phi(W_o^{[m]}, W_q^{[m]})|_m$ in Eq. (6). This improves the discriminative power of the features. Compared with Jia et al. (2014), the learned features in MMHIM are optimized for classification, thereby being able to represent visually similar actions. Methods in Oreifej and Liu (2013) and Yang et al. (2014) achieve surprisingly low performance compared with MMHIM due to noisy depth data. Thanks to the ability of effectively using visual data, MMHIM extracts discriminative features from visual data, and thus outperforms (Oreifej and Liu 2013; Yang et al. 2014). On the other hand, MMHIM represents features in a matrix form. This allows us to construct a low-rank bilinear model that can improve the generalization power. The learned features and parameter matrices in MMHIM are visualized in Fig. 8.

4.2.3 Single Modality

We evaluate the performance of MMHIM in single modality scenario, and investigate the contribution of each modality in recognition. In this experiment, we train MMHIM using one single modality data (RGB or depth), and test it using the same modality data. MMHIM and Structured MMHIM-1 are compared to linear SVM (with HOG, HOF, and HOGHOF features), bilinear SVM, and BHIM (Kong and Fu 2015).³ In addition, we independently build two MMHIMs, each of which is fed with RGB or depth modality. The prediction scores from the two MMHIMs are averaged to predict action labels. We refer to this method as MMHIM-ind. Results are shown in Table 3.

Recognition accuracy in Table 3 shows that MMHIM achieves superior performance in a majority of cases. Despite the method HOGHOF+linear SVM achieves better results over HOG or HOF-based method, these linear SVM methods do not fully take advantage of rich correlations between multiple features. MMHIM, on the contrary, finds better feature spaces to suppress noise and extract discriminative features, and thus achieves significantly better performance. MMHIM outperforms bilinear SVM by 7.20% given depth data. Compared with bilinear SVM, MMHIM utilizes information measure to compress data, and elegantly reduces redundancy in the data. MMHIM also outperforms BHIM due to the learning of discriminative private features. Given single RGB or depth modality, the “shared features”⁴ O and the private features $Q^{[m]}$ in MMHIM may capture similar discriminative information. The private features capture little additional discriminative information compared to the shared features, and thus the improvement of MMHIM over BHIM is only around 1%. The improvement of MMHIM over BHIM increases to 2.29% if two modalities are provided. The shared features will be compromised in order to capture the overlapping information between the two modalities. Consequently,

³ Please refer to the supplemental material for formulations of bilinear SVM, BHIM, and MMHIM in single modality learning

⁴ Technically, the feature O here is not shared between two modalities as it is only computed from RGB data.

Table 3 Comparison results on NEU-UB Dataset given depth-only, RGB-only, and RGB-D data

Methods	RGB (%)	Depth (%)	Depth+RGB (%)
HOG+linear SVM	76.67	46.14	72.11
HOF+linear SVM	77.19	36.49	69.47
HOGHOF+linear SVM	77.19	47.54	75.44
Bilinear SVM (Pirsiavash et al. 2009)	78.07	51.05	80.70
BHIM (Kong and Fu 2015)	78.77	56.84	81.22
MMHIM-ind	80.00	58.25	68.25
Structured MMHIM-1	79.82	61.05	82.63
MMHIM	80.00	58.25	83.51

it may not be able to capture all the discriminative information within one modality. In this case, private features will be acting as complementary features to shared features, and thus the improvement increases. Results also clearly show that RGB information contributes more to the classification than the depth information on this dataset. This is possibly because the motion variations of depth videos in different categories are subtle, and depth videos in NEU-UB dataset are much noisier than RGB videos.

Results also show that MMHIM significantly outperforms MMHIM-ind by 15.26% given RGB-D (Depth+RGB) data. MMHIM-ind achieves low performance because depth data are noisy and will lower down the overall performance if we simply combine prediction results on RGB and depth data. MMHIM-ind trains on RGB and depth data independently, and does not leverage the correlations between the two modalities to improve the discriminability. We observe in experiment that prediction results of MMHIM-ind on RGB data are always inconsistent with the predictions on noisy depth data. If we simply combine the prediction scores of the two modalities by averaging, the accurate results on RGB data will be compromised, and thus the overall performance is low (68.25% in Table 3). On the contrary, MMHIM learns shared features and private features from RGB-D data jointly. Even though the depth data are noisy, MMHIM projects the data onto a subspace where the discriminative information of depth data and RGB data can be jointly discovered. This enhances the discriminative power of the learned features from depth data, and thus makes MMHIM outperform MMHIM-ind.

4.2.4 Knowledge Transfer

We evaluate the performance of our MMHIM in missing modality scenarios, where one modality is missing in training or testing. MMHIM is tested in the following four scenarios: depth data are missing in testing (RGB-D \rightarrow RGB), RGB data are missing in testing (RGB-D \rightarrow Depth), depth data are missing in training (RGB \rightarrow RGB-D), and RGB

data are missing in training (Depth \rightarrow RGB-D). We compare MMHIM with linear SVM, bilinear SVM, BHIM (Kong and Fu 2015) and Structured MMHIM-1, and investigate how the knowledge transferred from observed modality influences the performance of the five methods. We also build a baseline, MMHIM-ind, which trains an independent classifier for each modality, and uses the relevant classifier in prediction phase according to the presence of modality.

Recognition results in Table 4 show that MMHIM significantly outperforms linear and bilinear SVM due to the use of a matrix form feature representation, and the learned cross-modal features and private features. MMHIM outperforms MMHIM-ind as the shared information between depth and RGB modalities is discovered and used in the training of MMHIM. The performance gap is significant when RGB modality is missing in testing (RGB-D \rightarrow Depth) as RGB data are more discriminative in the NEU-UB dataset. Without RGB modality in testing, MMHIM-ind will be confused by noisy depth data. The effectiveness of private features in MMHIM can be clearly seen from performance gap between MMHIM and BHIM, especially in the scenario of RGB-D \rightarrow Depth, where MMHIM outperforms BHIM by 22.11%. By extracting private features from the testing depth videos, MMHIM is capable of using extra discriminative information for classification, while BHIM can only use raw noisy depth features.

4.2.5 Sensitivity to Dimensionality

We also evaluate the sensitivity of MMHIM to the dimensionality d_o of the shared feature space and d_q of the private feature space. The performance variations of MMHIM given different parameters d_o and d_q are visualized in Fig. 9 in order to better show how MMHIM is affected by these two parameters. d_o and d_q range from 5 to 90 with step 5 (d_o or d_q should be less than $n_f = 93$). Results show that MMHIM is not sensitive to the dimensionality d_o and d_q . The performance variation with respect to d_o is only 1.4% and it is only 1.75% with respect to d_q .

Table 4 Knowledge transfer results on NEU-UB dataset

Methods	RGB-D → RGB (%)	RGB-D → Depth (%)	RGB → RGB-D (%)	Depth → RGB-D (%)
Linear SVM	16.67	36.32	63.85	47.92
MMHIM-ind	80.00	58.25	N/A	N/A
Bilinear SVM (Pirsiavash et al. 2009)	68.77	58.60	57.72	50.53
BHIM (Kong and Fu 2015)	81.93	59.82	77.19	54.91
Structured MMHIM-1	80.70	82.46	76.84	55.26
MMHIM	81.93	81.93	78.60	55.26

$X \rightarrow Y$ denotes that X is the training data and Y is the testing data. $d_o = d_q = 10$ for both bilinear SVM and MMHIM, and dimensionality of features in linear SVM is $n_{xyt} \cdot n_f$. “N/A” indicates the model is not applicable

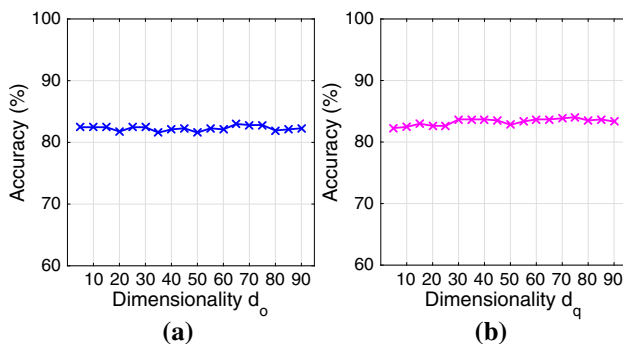


Fig. 9 RGB-D action recognition results on our MMHIM on NEU-UB dataset with different **a** dimensionality d_o of the shared feature space and **b** dimensionality d_q of the private feature space

4.3 MSR Action Pairs Dataset

We follow the same preprocessing scheme used in the previous dataset. The same split training strategy in Oreifej and Liu (2013) is adopted, where RGB-D videos of the first 5 subjects are used as training data.

4.3.1 Comparison Experiment

We compare with existing methods (Oreifej and Liu 2013; Wang et al. 2012b; Yang et al. 2012, 2014; Jia et al. 2014; Kong and Fu 2015), and use linear SVM as baseline. We also extend the bilinear SVM (Pirsiavash et al. 2009) to capture two-modality data, and use it as baseline.

Results in Table 5 show that MMHIM outperforms all the comparison approaches. MMHIM achieves 100% accuracy, significantly outperforming linear SVM by 12.22%. Linear SVM simply concatenates RGB and depth features into a long vector, while MMHIM finds the optimal space for fusing the two features. Both bilinear SVM and MMHIM learn a shared feature space for feature fusion. However, MMHIM gains 3.89% improvement over bilinear SVM, demonstrating the benefit of noise reduction in MMHIM. In order to reduce noise in data projection, MMHIM uses the informa-

Table 5 Recognition accuracy of comparison methods on MSR Action Pairs dataset

Methods	Accuracy (%)
Linear SVM	87.78
Bilinear SVM (Pirsiavash et al. 2009)	96.11
Deep Motion Maps (Yang et al. 2012)	66.11
Skeleton+LOP+Pyramid (Wang et al. 2012b)	82.22
LTTL (Jia et al. 2014)	91.48
HON4D (Oreifej and Liu 2013)	96.67
SNV (Yang et al. 2014)	98.89
BHIM (Kong and Fu 2015)	100
Structured MMHIM-1	97.78
Structured MMHIM-2	100
MMHIM	100

Bold values indicate the best result in comparison

tion measure $\phi(W_o^{[m]}, W_q^{[m]})|_m$ in Eq. (6) to compress data and reduce information irrelevant to our recognition task. MMHIM also outperforms (Oreifej and Liu 2013; Wang et al. 2012b; Yang et al. 2012, 2014), which shows the benefits of effectively utilizing both visual and depth data, and representing features in a matrix form. Using a matrix form feature representation allows us to construct a low-rank bilinear model that can improve the generalization power. The effectiveness of a matrix form feature representation can also be clearly seen from the performance difference between MMHIM and Structured MMHIM-1. The learned features and parameter matrices in MMHIM are visualized in Fig. 10.

4.3.2 Sensitivity to Parameters

In this experiment, we investigate the sensitivity of MMHIM to the maximum rank of d_o , the parameter η and the parameter λ in Eq. (2). d_q is fixed to 10.

We first test the sensitivity of MMHIM to the maximum rank d_o . MMHIM is compared with linear SVM and bilinear SVM with various d_o values. Note that there are a total of $n_{xyt} \cdot d_o$ elements in the shared space for each modality in

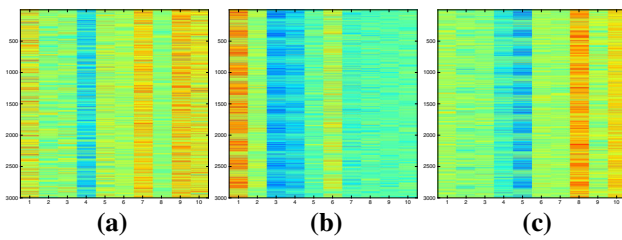


Fig. 10 Visualizations of **a** the projected visual features $X^{[v]}W_q^{[v]}$, **b** the projected depth features $X^{[z]}W_q^{[z]}$ in the private space, and **c** the learned cross-modal features O in the shared space

MMHIM and bilinear SVM. To conduct a fair comparison, for linear SVM, we use PCA to reduce the dimensionality of feature vectors of each modality to $n_{xyt} \cdot d_o$, making sure all the five methods have the same number of elements in the low-dimensional features. The projected visual and depth features are concatenated into a long vector and fed to linear SVM. In bilinear SVM and MMHIM, the original feature matrix $X^{[m]}$ is projected by $W_o^{[m]}$. The rank parameter d_o is set to 1, 5, and 31, respectively.

The performance of the five methods on depth features, RGB features, and RGB-D features are shown in Table 6. Results indicate that MMHIM achieves higher performance in most of the cases given low-dimensional features, and its performance on RGB-D data is not sensitive to parameter d_o . When $d_o = 1$, the projected feature matrices may lose certain amount of information. However, the structural information is reserved in MMHIM, resulting in significant higher performance over linear SVM. In addition, the learned shared space in MMHIM is optimized for classification, while it is not the case in PCA. Compared with bilinear SVM, noisy information is reduced in MMHIM, and thus it achieves superior performance, for example, an increase of 26.66% given RGB-D data.

Even though linear SVM captures full information from visual and depth features when $d_o = 31$, it simply concatenates depth and RGB features, suggesting that the similarities between the two types of features are directly compared. This

may not be appropriate since they are from different distributions. In contrast, our MMHIM solves this problem by projecting the two features into a shared feature space. The learning of a shared feature space allows us to effectively use the two types of features for recognition. MMHIM outperforms bilinear SVM especially in the case of RGB data, where the performance gain is 18.89%.

MMHIM achieves superior results over BHIM when $d_o = 5$ due to the learning of private features. An improvement of 8.33% can be observed given RGB data. The learned cross-modal features in BHIM lose too much discriminative information using the information measure $\phi(W_o^{[m]}, W_q^{[m]})$. MMHIM, on the contrary, uses extra private features to capture rich discriminative information.

Figure 11 illustrates RGB-D action recognition performance variations of MMHIM on MSR Action Pairs dataset given different values of parameter λ and parameter η . Parameter λ is set: $\lambda \in \{0.00001, 0.00005, 0.0001, 0.0005, 0.001, 0.005, 0.01, 0.05, 0.1, 0.5, 1, 5, 10, 50, 100\}$, and parameter η is set: $\eta \in \{0.0001, 0.0005, 0.001, 0.005, 0.01, 0.05, 0.1, 0.5, 1\}$. Figure 11a shows that MMHIM is insensitive to λ if $\lambda \leq 50$; the largest performance variations is only 2.22%. However, the performance drops to 77.78% if $\lambda = 100$ as MMHIM focuses on the margin regularizer but pays less attention to the training loss. Figure 11b demonstrates that MMHIM is insensitive to parameter η when $\eta \leq 0.1$; the variation is only 2.22%. The performance drops to 86.11% when $\eta = 0.5$ and it further drops to 67.22% when $\eta = 1$. This is because if η is set to some relatively large values ($\eta = 1$ for example), MMHIM prefers to learn very complex decision boundaries to minimize training loss. This affects the generalization performance of MMHIM, and results in low recognition performance.

4.3.3 Knowledge Transfer

We evaluate the performance of our MMHIM, and investigate the effectiveness of the cross-modal features, the private features, and the information measure if one modal-

Table 6 Comparison results with various dimensionality d_o of the feature space

Methods	$d_o = 1$			$d_o = 5$			$d_o = 31$		
	Depth (%)	RGB (%)	RGB-D (%)	Depth (%)	RGB (%)	RGB-D (%)	Depth (%)	RGB (%)	RGB-D (%)
Linear SVM	47.22	42.78	51.67	72.78	70.00	75.00	86.11	87.22	87.78
Bilinear SVM (Pirsiavash et al. 2009)	53.89	50.00	70.56	90.00	87.22	91.11	92.78	80.00	96.11
BHIM (Kong and Fu 2015)	83.33	91.11	96.11	88.33	76.11	98.33	93.89	97.22	100
Structured MMHIM-1	83.33	91.11	97.22	90.00	82.78	98.33	95.00	98.33	100
MMHIM	84.44	91.67	97.22	90.00	84.44	100	96.11	98.89	100

The dimensionality of features for each modality in linear SVM is $n_{xyt} \cdot d_o$

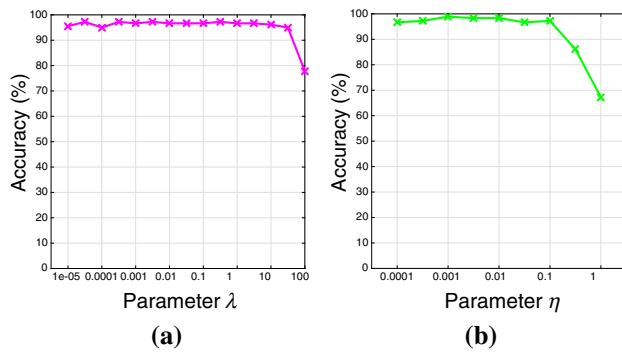


Fig. 11 RGB-D action recognition results of our MMHIM on MSR Action Pairs dataset with different values of **a** parameter λ and **b** parameter η

ity is missing in training or testing. MMHIM is tested in four scenarios as in the previous dataset: RGB-D \rightarrow RGB, RGB-D \rightarrow Depth, RGB \rightarrow RGB-D, and Depth \rightarrow RGB-D. We compare MMHIM with linear SVM, bilinear SVM, and BHIM (Kong and Fu 2015), and verify their performance in knowledge transfer.

Recognition results in Table 7 show that MMHIM markedly outperforms BHIM, linear and bilinear SVM in this knowledge transfer experiment. MMHIM achieves better performance over BHIM in this experiment, especially when RGB data are missing, i.e. RGB-D \rightarrow Depth and Depth \rightarrow RGB-D. The underlying reason is that private features can capture extra discriminative information for classification. If RGB modality is missing, the depth modality itself is noisy and can affect the performance. Using private features, the depth modality is projected to two spaces (one for learning shared features and the other for private features), and thus extra discriminative information can be used for classification. The performance gain over linear SVM is noticeable, indicating the effectiveness of using a matrix form feature representation, and cross-modal and private features in MMHIM. The improvement of MMHIM over bilinear SVM is also significant, especially in RGB \rightarrow RGB-D where the increase is 17.22%. MMHIM achieves better performance over BHIM in this experiment, especially when RGB data are missing, i.e. RGB-D \rightarrow Depth and Depth \rightarrow RGB-D.

4.4 MSR Daily Activity Dataset

RGB and depth sequences in this dataset are spatially and temporally normalized, and the people of interest are extracted from these sequences. We follow the same training protocol in Wang et al. (2012b). Skeleton feature vectors in Sect. 3.3 are adopted to build a MMHIM+Skeleton method, where skeleton features are extracted using Du et al. (2015). Our methods are first compared with existing approaches (Yang et al. 2012; Liu and Shao 2013; Zanfir et al. 2013; Oreifej and Liu 2013; Wang et al. 2012b; Yang et al. 2014; Kong and Fu 2015; Hu et al. 2015) on this dataset, and then evaluated given RGB, depth, and RGB-D data, respectively. Linear SVM and bilinear SVM are used as baseline.

Comparison results are shown in Table 8. MMHIM+skeleton achieves superior performance over state-of-the-art methods due to the effective use of visual, depth, and skeleton data. Its performance gain over MMHIM (without skeleton data) is 8.10%, suggesting the benefit of using skeleton feature vectors. MMHIM+skeleton significantly outperforms BHIM due to the fusion of three types of features and the ability of using private features. Structured MMHIM-2 outperforms MMHIM by 1.88% on MSR Daily Activity dataset, indicating the importance of modeling correlations between frames on this dataset. MMHIM notably outperforms linear SVM by 22.50% owing to the learning of shared and private feature spaces for the two types of features, and a matrix form representation that naturally captures spatiotemporal structural information. Recognition accuracy of MMHIM is also higher than bilinear SVM thanks to the use of information measure, which is helpful in suppressing useless information and noise. MMHIM outperforms recent surface normal-based approaches (Oreifej and Liu 2013; Yang et al. 2014). Although these approaches essentially capture structural information in the feature design stage, they only focus on depth sequences, and do not utilize valuable visual information. In addition, the two approaches use the full length feature vectors and do not learn a better feature space for classification. MMHIM achieves better performance than the actionlet ensemble approach (Wang et al. 2012b) since we

Table 7 Knowledge transfer results on MSR Action Pairs dataset

Methods	RGB-D \rightarrow RGB (%)	RGB-D \rightarrow Depth (%)	RGB \rightarrow RGB-D (%)	Depth \rightarrow RGB-D (%)
linear SVM	83.33	81.67	87.22	86.11
Bilinear SVM (Pirsiavash et al. 2009)	90.56	93.89	81.67	91.67
BHIM (Kong and Fu 2015)	97.78	92.78	97.78	93.33
Structured MMHIM-1	97.78	93.89	97.78	95.00
MMHIM	98.89	94.44	98.89	96.11

$X \rightarrow Y$ denotes that X is the training data and Y is the testing data. $d_o = 31$ for both bilinear SVM and MMHIM, and dimensionality of features in linear SVM is $n_{xyt} \cdot d_o$. The number of elements in the input feature vector/matrix to the three methods is the same

Table 8 Recognition accuracy of comparison methods on MSR Daily Activity Dataset

Methods	Skeleton	Accuracy (%)
Linear SVM	No	65.00
Bilinear SVM (Pirsiavash et al. 2009)	No	85.63
Depth Motion Maps (Yang et al. 2012)	No	43.13
Moving Pose (Zanfir et al. 2013)	Yes	73.80
CNN (Liu and Shao 2013)	No	75.00
DBN (Liu and Shao 2013)	No	77.90
Local HON4D (Oreifej and Liu 2013)	No	80.00
RGGP (Liu and Shao 2013)	No	82.10
Actionlet Ensemble (Wang et al. 2012b)	Yes	85.75
SNV (Yang et al. 2014)	No	86.25
BHIM (Kong and Fu 2015)	No	86.88
JOULE (Hu et al. 2015)	Yes	95.00
Structured MMHIM-1	No	86.88
MMHIM	No	87.50
Structured MMHIM-2	No	89.38
MMHIM+Skeleton	Yes	95.60

Bold value indicates the best result in comparison

Table 9 Comparison results on MSR Daily Activity Dataset given depth-only, RGB-only, and RGB-D data

Methods	Depth (%)	RGB (%)	Depth+RGB (%)
HOG+linear SVM	70.00	55.56	66.67
HOF+linear SVM	67.78	57.78	68.89
HOGHOF+linear SVM	67.78	57.78	71.11
Bilinear SVM (Pirsiavash et al. 2009)	72.50	67.50	85.63
BHIM (Kong and Fu 2015)	81.88	77.50	86.88
MMHIM	82.50	79.38	87.50

elegantly use visual and depth information, and effectively compress informative cues and remove noise before classification. We also compare with two deep models, CNN and DBN in Liu and Shao (2013). The two models do not perform very well (even worse than the hand-crafted HON4D method with linear SVM classifier, and bilinear SVM method) possibly due to small size of the training data. This also demonstrates that linear and bilinear models are able to learn complex mappings compared to deep models, especially if the training data is scarce.

Performance of the proposed MMHIM on the RGB-only, depth-only, and RGB-D data in the MSR Daily Activity dataset is also reported in this paper. Recognition accuracy in Table 9 shows that MMHIM achieves satisfactory results even though only one modality of features is provided. MMHIM outperforms BHIM by 1.88% given RGB data due to the use of private features. RGB cue is not as discriminative as depth cue in the MSR Daily Activity dataset as the motions in various actions are similar and objects are relatively small. However, using private features can capture additional discriminative information that may be lost in the learning of share features. The

improvement of MMHIM over linear and bilinear SVM is remarkable, which is 14.72 and 10.00% given depth data, respectively. This mainly contributes to the use of private features and the information measure in MMHIM. Private features enable us to use extra discriminative information, and information measure reduces non-helpful information for classification.

4.5 Discussions

The three datasets, NEU-UB Action dataset, MSR Action Pairs dataset, MSR Daily Activity dataset, vary in terms of actions and noise. NEU-UB dataset contains 6 actions with basic limb movement, including “lateral bend”, “left leg lift”, “right leg lift”, “squat”, “step backward”, “step forward”. Actions in MSR Action Pairs dataset are human-object interactions, such as “pick up/put down box”, “lift/place box”, “push/pull chair”, “wear/take off hat”, “put on/off backpack”, “stick/remove poster”. MSR Daily Activity dataset consists of actions that are often seen in daily life, such as “drink”, “call cellphone”, and “lay down on sofa”.

The performance variation on the three datasets is mainly due to the noise level in depth videos, and visual similarity and motion-object relationships in RGB videos. Compared to existing MSR Action Pairs dataset and MSR Daily Activity dataset, depth videos in NEU-UB dataset are more noisy. This can be clearly seen from the performance difference using depth videos on these datasets. Given depth modality in training and testing, MMHIM only achieves the recognition accuracy of 58.25% on NEU-UB dataset, 24.25% lower than the results on MSR Daily Activity dataset.

Some of RGB videos of different categories in NEU-UB dataset are visually similar, for example, “step forward” and “step backward”. The major difference between them is the temporal size change of the human body in videos, which is difficult to capture. Therefore, MMHIM only achieves 80.00% accuracy in RGB modality learning. By comparison, the major difference in actions in MSR Action Pairs dataset is the motion-object shape relationship. The relationship is easy to capture as particular human motion is acted to operate an object, and objects differ in their sizes. That is why a majority of approaches achieve high performance on this dataset. RGB action videos in MSR Daily Activity dataset usually involve objects of small size, such as a cup, a book, and a cellphone. This makes the motion-object relationships difficult to capture in actions involving small objects, and MMHIM achieves the lowest performance (79.38%) among the three datasets.

MMHIM is implemented in Matlab and trained on a 3.5 GHz machine with 64 GB memory. The training on NEU-UB dataset, MSR Action Pairs dataset, and MSR Daily Activity dataset takes about 4, 7, and 6 min, respectively. Memory usage for training on the three datasets are 3, 2, and 1.5 GB, respectively.

5 Conclusion

We have proposed a max-margin heterogeneous information machine (MMHIM) for action recognition from RGB-D sequences. MMHIM effectively utilizes RGB visual and depth data for classification. The two modality data are used to learn cross-modality features to capture their correlations, and learn private features to capture their modality specific characteristics. Features are represented in a matrix form in MMHIM, which essentially characterizes spatiotemporal relationships of human body parts. This naturally allows us to construct a low-rank bilinear model to use these feature matrices, and learn cross-modality and private features. These two types of features are denoised using the information measure. Classification is performed using the denoised shared and private features. We learn a low-rank MMHIM by directly minimizing the rank of the model, in order to increase the generalization power. An efficient optimization algorithm

is proposed in this work with an off-the-shelf SVM solver as the inner optimization solver. The MMHIM is extensively evaluated on three public RGB-D action datasets, and outperforms state-of-the-art approaches.

Acknowledgements This work is supported in part by the NSF IIS Award 1651902, NSF CNS Award 1314484, ONR Award N00014-12-1-1028, ONR Young Investigator Award N00014-14-1-0484, and U.S. Army Research Office Young Investigator Award W911NF-14-1-0218.

References

- Andrew, G., Arora, R., Bilmes, J., & Livescu, K. (2013). Deep canonical correlation analysis. In *ICML*.
- Argyriou, A., Evgeniou, T., & Pontil, M. (2008). Convex multi-task feature learning. In *IJCV*.
- Bo, L., Lai, K., Ren, X., & Fox, D. (2011). Object recognition with hierarchical kernel descriptors. In *CVPR*.
- Chen, L., Li, W., & Xu, D. (2014). Recognizing RGB images by learning from RGB-D data. In *CVPR*.
- Dalal, N., & Triggs, B. (2005). Histograms of oriented gradients for human detection. In *IEEE computer society conference on CVPR 2005* (Vol. 1, pp. 886–893). doi:10.1109/CVPR.2005.177.
- Do, T. M. T., & Artieres, T. (2009). Large margin training for hidden markov models with partially observed states. In *ICML*.
- Dollar, P., Rabaud, V., Cottrell, G., & Belongie, S. (2005). Behavior recognition via sparse spatio-temporal features. In *ICCV VS-PETS*.
- Du, Y., Wang, W., & Wang, L. (2015). Hierarchical recurrent neural network for skeleton based action recognition. In *CVPR*.
- El, R. O., Rosman, G., Wetzler, A., Kimmel, R., & Bruckstein, A. M. (2015). Rgbd-fusion: Real-time high precision depth recovery. In *CVPR*.
- Fernando, B., Anderson, P., Hutter, M., & Gould, S. (2016). Discriminative hierarchical rank pooling for activity recognition. In *CVPR*.
- Fernando, B., Gavves, E., Ghodrati, J. O. M. A., & Tuytelaars, T. (2015). Modeling video evolution for action recognition. In *CVPR*.
- Hadfield, S., & Bowden, R. (2013). Hollywood 3D: Recognizing actions in 3D natural scenes. In *CVPR*.
- Hu, J. F., Zheng, W. S., Lai, J., & Zhang, J. (2015). Jointly learning heterogeneous features for rgb-d activity recognition. In *CVPR*.
- Ji, S., Xu, W., Yang, M., & Yu, K. (2013). 3D convolutional neural networks for human action recognition. In *PAMI*.
- Jia, C., Kong, Y., Ding, Z., & Fu, Y. (2014). Latent tensor transfer learning for RGB-D action recognition. In *ACM Multimedia*.
- Joachims, T., Finley, T., & Yu, C. N. (2009). Cutting-plane training of structural SVMs. *Machine Learning*, 77(1), 27–59.
- Karpathy, A., Toderici, G., Shetty, S., Leung, T., & Sukthankar, R., Fei-Fei, L. (2014). Large-scale video classification with convolutional neural networks. In *CVPR*.
- Klaser, A., Marszalek, M., & Schmid, C. (2008). A spatio-temporal descriptor based on 3d-gradients. In *BMVC* (pp. 1–10).
- Kobayashi, T. (2014). Low-rank bilinear classification: Efficient convex optimization and extensions. In *IJCV*.
- Kong, Y., & Fu, Y. (2015). Bilinear heterogeneous information machine for rgb-d action recognition. In *CVPR*.
- Kong, Y., Jia, Y., & Fu, Y. (2014). Interactive phrases: Semantic descriptions for human interaction recognition. In *PAMI*.
- Kong, Y., Kit, D., & Fu, Y. (2014). A discriminative model with multiple temporal scales for action prediction. In *ECCV*.
- Koppula, H.S., & Saxena, A. (2013). Learning spatio-temporal structure from RGB-D videos for human activity detection and anticipation. In *ICML*.

- Lan, T., Wang, Y., Yang, W., Robinovitch, S. N., & Mori, G. (2012). Discriminative latent models for recognizing contextual group activities. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *34*(8), 1549–1562.
- Laptev, I. (2005). On space-time interest points. *International Journal of Computer Vision*, *64*(2/3), 107–123.
- Li, W., Zhang, Z., & Liu, Z. (2010). Action recognition based on a bag of 3D points. In *CVPR workshop*.
- Lin, Y. Y., Hua, J. H., Tang, N. C., Chen, M. H., & Liao, H. Y. M. (2014). Depth and skeleton associated action recognition without online accessible RGB-D cameras. In *CVPR*.
- Liu, J., Ali, S., & Shah, M. (2008). Recognizing human actions using multiple features. In *CVPR* (pp. 1–8).
- Liu, J., Kuipers, B., & Savarese, S. (2011). Recognizing human actions by attributes. In *CVPR* (pp. 3337–3344).
- Liu, L., & Shao, L. (2013). Learning discriminative representations from RGB-D video data. In *IJCAI*.
- Lu, C., Jia, J., & Tang, C. K. (2014). Range-sample depth feature for action recognition. In *CVPR*.
- Luo, J., Wang, W., & Qi, H. (2013). Group sparsity and geometry constrained dictionary learning for action recognition from depth maps. In *ICCV*.
- Ma, S., Sigal, L., & Sclaroff, S. (2016). Learning activity progression in lstms for activity detection and early detection. In *CVPR*.
- Marszałek, M., Laptev, I., & Schmid, C. (2009). Actions in context. In *Proceedings of IEEE conference on computer vision and pattern recognition*.
- Ngiam, J., Khosla, A., Kim, M., Nam, J., Lee, H., & Ng, A. Y. (2011). Multimodal deep learning. In *ICML*.
- Ni, B., Moulin, P., Yang, X., & Yan, S. (2015). Motion part regularization: Improving action recognition via trajectory group selection. In *CVPR*.
- Ni, B., Wang, G., & Moulin, P. (2011). RGBD-HuDaAct: A color-depth video database for human daily activity recognition. In *ICCV Workshop on CDC3CV*.
- Ofli, F., Chaudhry, R., Kurillo, G., Vidal, R., & Bajcsy, R. (2013). Berkeley MHAD: A comprehensive multimodal human action database. In *Proceedings of the IEEE Workshop on Applications on Computer Vision*.
- Oreifej, O., & Liu, Z. (2013). HON4D: Histogram of oriented 4D normals for activity recognition from depth sequences. In *CVPR*. doi:10.1109/CVPR.2013.98.
- Pirsiavash, H., Ramanan, D., & Fowlkes, C. (2009). Bilinear classifiers for visual recognition. In *NIPS*.
- Raptis, M., & Sigal, L. (2013). Poselet key-framing: A model for human activity recognition. In *CVPR*.
- Raptis, M., & Soatto, S. (2010). Tracklet descriptors for action modeling and video analysis. In *ECCV*.
- Ryoo, M., & Aggarwal, J. (2009). Spatio-temporal relationship match: Video structure comparison for recognition of complex human activities. In *ICCV* (pp. 1593–1600).
- Schüldt, C., Laptev, I., & Caputo, B. (2004). Recognizing human actions: A local SVM approach. In *ICPR*.
- Shotton, J., Girshick, R., Fitzgibbon, A., Sharp, T., Cook, M., Finocchio, M., Moore, R., Kohli, P., Criminisi, A., Kipman, A., & Blake, A. (2013). Efficient human pose estimation from single depth images. In *PAMI*.
- Simonyan, K., & Zisserman, A. (2014). two-stream convolutional networks for action recognition in videos. In *NIPS*.
- Srivastava, N., & Salakhutdinov, R. (2014). Multimodal learning with deep boltzmann machines. *JMLR*, *15*, 2949–2980.
- Sung, J., Ponce, C., Selman, B., & Saxena, A. (2012). Unstructured human activity detection from rgb-d images. In *ICRA*.
- Tang, K., Fei-Fei, L., & Koller, D. (2012). Learning latent temporal structure for complex event detection. In *CVPR*.
- Tenenbaum, J. B., & Freeman, W. T. (2000). Separating style and content with bilinear models. *Neural Computation*.
- Teo, C.H., Le, Q., Smola, A., & Vishwanathan, S. (2007). A scalable modular convex solver for regularized risk minimization. In *KDD*.
- Tishby, N., Pereira, F. C., & Bialek, W. (1999). The information bottleneck method. In *Proceedings of the 37-th annual allerton conference on communication, control and computing*, pp. 368–377.
- Vondrick, C., Pirsiavash, H., & Torralba, A. (2016). Anticipating visual representations from unlabeled video. In *CVPR*.
- Wang, H., & Schmid, C. (2013). Action recognition with improved trajectories. In *IEEE International Conference on Computer Vision*, Sydney, Australia. <http://hal.inria.fr/hal-00873267>
- Wang, J., Liu, Z., Chorowski, J., Chen, Z., & Wu, Y. (2012a). Robust 3D action recognition with random occupancy patterns. In *ECCV* (pp. 872–885).
- Wang, J., Liu, Z., Wu, Y., & Yuan, J. (2012b). Mining actionlet ensemble for action recognition with depth cameras. In *CVPR*.
- Wang, P., Shen, X., Lin, Z., Cohen, S., Price, B., & Yuille, A. L. (2015). Towards unified depth and semantic prediction from a single image. In *CVPR*.
- Wang, W., Arora, R., Livescu, K., & Bilmes, J. (2015). On deep multi-view representation learning. In *ICML*.
- Wolf, L., Jhuang, H., & Hazan, T. (2007). Modeling appearances with low-rank svm. In *CVPR*.
- Wu, C., Zhang, J., Savarese, S., & Saxena, A. (2015). Watch-n-patch: Unsupervised understanding of actions and relations. In *CVPR*.
- Xia, L., & Aggarwal, J. (2013). Spatio-temporal depth cuboid similarity feature for activity recognition using depth camera. In *CVPR*.
- Xie, P., & Xing, E. P. (2013). Multi-modal distance metric learning. In *IJCAI*.
- Xu, C., & Cheng, L. (2013). Efficient hand pose estimation from a single depth image. In *ICCV*.
- Xu, C., Tao, D., & Xu, C. (2014). Large-margin multi-view information bottleneck. *PAMI*, *36*(8), 1559–1572.
- Yang, X., & Tian, Y. (2014). Super normal vector for activity recognition using depth sequences. In *CVPR*.
- Yang, X., Zhang, C., & Tian, Y. (2012). Recognizing actions using depth motion maps-based histograms of oriented gradients. In *ACM Multimedia*. doi:10.1145/2393347.2396382.
- Zanfir, M., Leordeanu, M., & Sminchisescu, C. (2013). The moving pose: An efficient 3D kinematics descriptor for low-latency action recognition and detection. In *ICCV*.
- Zhang, J., Kan, C., Schwing, A. G., & Urtasun, R. (2013). Estimating the 3d layout of indoor scenes and its clutter from depth sensors. In *ICCV*.
- Zhou, Y., Ni, B., Hong, R., Wang, M., & Tian, Q. (2015). Interaction part mining: A mid-level approach for fine-grained action recognition. In *CVPR*.