Bilinear Heterogeneous Information Machine for RGB-D Action Recognition

Yu Kong[1] and Yun Fu[1,2]
[1]Department of Electrical and Computer Engineering,
[2]College of Computer and Information Science.
Northeastern University, Boston, MA, USA
{yukong,yunfu}@ece.neu.edu

Abstract

This paper proposes a novel approach to action recognition from RGB-D cameras, in which depth features and RGB visual features are jointly used. Rich heterogeneous RGB and depth data are effectively compressed and projected to a learned shared space, 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. The proposed method is extensively evaluated on two public RGB-D action datasets, and achieves state-of-the-art results. It also shows promising results if RGB or depth data are missing in training or testing procedure.

1. Introduction

Action recognition from RGB-D cameras has been receiving increasing interests in the computer vision community due to the recent advance of easy-to-use and low-cost depth sensors such as Kinect sensors [16]. In addition to RGB visual data captured by conventional RGB cameras, depth data are provided in RGB-D cameras, encoding rich 3D structural information of the entire scene. Previous work [16, 13, 21, 5] showed that effective usage of 3D structural information facilitates recognition tasks as it simplifies intra-class motion variations and removes cluttered background noise.

Despite its effectiveness, those methods are only applicable when depth data are available. Methods developed in [25, 13, 23, 5] are particularly designed for depth data, and thus would fail if depth data are unavailable or missing in RGB-D cameras. In addition, depth data are noisy due to spatiotemporal discontinuous regions. This hinders the application of feature extraction methods such as surface normal [25, 13] and spatiotemporal interest points [23, 5] 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 in case of having depth information as a single cue.

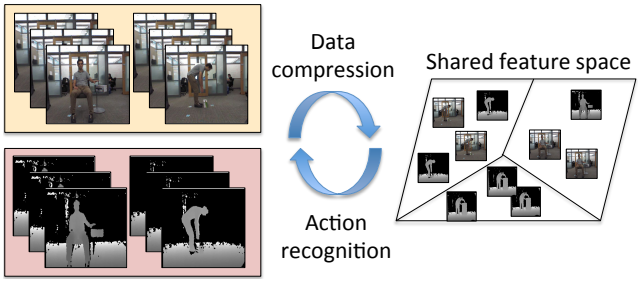


Figure 1. Our method projects and compresses both RGB visual features and depth features to a learned shared feature space. Classification boundaries are learned in the shared space for action recognition. This process iterates until convergence.

RGB data and depth data can be complementary to each other if one of them is missing. Implicit correlations between them can be learned to handle the case that 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 [25, 13] attempted to capture spatiotemporal correlation information of body part movements by aggregating features from neighborhoods. However, this information would unfortunately collapse as co-occurrence features

are concatenated into high dimensional vectors [18].

In this paper, we propose a novel bilinear heterogeneous information machine (BHIM) for action recognition from RGB-D sequences. BHIM learns cross-modal features that effectively capture heterogeneous visual and depth information. RGB and depth data are treated as two modalities in this work. We project the original features of the two modalities onto a shared space, and learn cross-modal features 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. Moreover, the features are "filtered" for noise removal in the projection procedure. We show in the experiment that the learned cross-modal 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 a classification parameter matrix, 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, which is the optimal space for building robust and effective cross-modal features for recognition. An information measure is incorporated in the learning of projection matrices to help to reduce noise in feature projection procedure. Classification is performed using the learned cross-modal features. The rank of the model is minimized from the viewpoint of generalization power and computational cost [22].

We propose an efficient algorithm to optimize BHIM. 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 an action classifier 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 [19, 4] being the solver for the inner optimization problem.

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

2. Related Work

Previous action recognition approaches mainly focus on RGB action videos [9, 15, 17, 6]. These studies used low-level interest point features [17], mid-level semantic features [9] or human pose [15], or learned features using deep learning technique [6]. However, misclassification exists due to large intra-class variations such as motion and pose.

Due to the advent of low-cost Kinect sensors [16], lots of attempts have been devoted to object recognition [3, 2] and action recognition [10, 13, 25, 5] from depth images. One of the main advantages of depth data is that they capture 3D structural information, which helps reduce background noise, and simplifies intra-class variations. Effective features have been proposed for recognition from depth data, such as action graph [10], histogram of oriented 4D normals [13], super normal vector [25], 4D interest point-based method [5], and depth spatiotemporal interest points [23]. Features from depth sequences can be encoded by [12], or be used to build actionlets [21] for recognition. Recent work [11] also showed that features of RGB-D data can also be learned using popular deep learning techniques.

Those methods only use depth data, and thus would fail if depth data are missing. In contrast, our method uses both RGB and depth data, and can handle the case if one modality is missing. Moreover, they use features in a vector form, in which spatiotemporal structures would easily collapse [18, 8]. In this work, we propose to use features in a matrix form, which naturally captures both spatiotemporal structural information and motion information. We show in the experiment that features in a matrix format significantly improve the performance even though the rank of the parameter matrices in BHIM is constrained to be 1.

Feature learning methods [8, 14, 1, 24] have been proposed to learn better feature representations for recognition. Different from [14], we elegantly use features from two modalities for recognition. In contrast to [8], we use the Frobenius norm instead of the trace norm, which allows us to use existing efficient SVM solvers. In addition, we use an effective information measure to produce more compact cross-modal features, while this was not considered in [14, 8]. Method [24] extends information bottleneck [20] to a multi-view model. In contrast to their work, we learn a low-rank bilinear model, which shows better generalization power than a linear model. In addition, our method can rec-

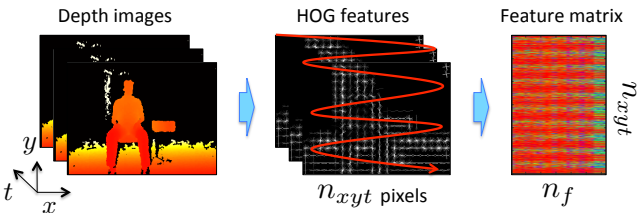


Figure 2. Feature matrix of size $n_{xyt} \times n_f$ is constructed from features (e.g., HOG) computed on all the frames. n_{xyt} is the total number of pixels in all the feature frames, and n_f is the dimensionality of each local feature.

ognize actions if one modality is missing. However, those methods were not designed for handling missing modality and their performance is not clear.

3. Bilinear Heterogeneous Information Machine

The goal of this work is to utilize heterogeneous features from RGB-D action videos, and learn shared cross-modal 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. Note that $X_i^{[v]}$ and $X_i^{[z]}$ in our work are defined as feature matrices of size $n_{xyt} \times n_f$, different from feature vectors containing $n_{xyt} \times n_f$ elements that are popularly used in computer vision community. In this work, features $X_i^{[v]}$ and $X_i^{[z]}$ (such as histogram of oriented gradient) 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. 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 P_v and P_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: $P_v : \mathcal{X}_v \to \mathcal{O}$, and $P_z : \mathcal{X}_z \to \mathcal{O}$. After learning the projection functions, a classification model G can be learned to classify actions given features in the shared space: $G : \mathcal{O} \to \mathcal{Y}$,

Instead of learning the projection functions, P_v and P_z, and the classification function G independently, we are interested in learning these functions simultaneously. Therefore, the learned projections are optimized for classification. We focus on learning a discriminant function $F : \mathcal{X} \times \mathcal{Y} \to \mathcal{R}$ that scores each training sample (X_i, y_i). The function F is applied to compute the compatibility between original RGB-D features X_i and the learned cross-modal features O_i, and between the features O_i and an action label y_i.

3.1. Model Formulation

Suppose we are given M ($M = 2$ in this work) types of modalities $X_i^{[m]}|_{m=1}^M$. Here, m is the index of modality, which can be either visual ($m = 1$) or depth ($m = 2$). We represent both of the two modality features in a matrix form in order to keep inherent spatiotemporal structure. In this paper, we are interested in a binary bilinear discriminant function $F(X_i, y|W) = \mathrm{Tr}(W^{\mathrm{T}} X_i) = \sum_{m=1}^M \mathrm{Tr}(W^{[m]\mathrm{T}} X_i^{[m]})$, which is a family of bilinear functions parameterized by a model weight matrix W. The one-vs-one scheme is adopted to extend our binary classifier to a multi-class classifier. One of the challenges in RGB-D action recognition is that the two modalities, RGB features and depth features, are in different feature spaces, and thus their similarities cannot be directly computed. We solve this problem by decomposing the parameter matrix $W^{[m]}$ for each modality into two components, $W_f^{[m]}$ and W_w: $W^{[m]} = W_w W_f^{[m]\mathrm{T}}$ (see Figure 3). Parameter matrix $W_f^{[m]} \in \mathcal{R}^{n_f \times d}$ ($m = 1, \cdots, M$) projects the m modality data, $X^{[m]}$, onto a learned shared space, and parameter matrix $W_w \in \mathcal{R}^{n_{xyt} \times d}$ is applied to classify the projected data regardless of modalities. W_w is a spatiotemporal template defined over d features at each spatiotemporal location. Obviously, the rank of the model parameter matrix $W^{[m]}$ will be enforced to be at most d.

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

$$y_i^* = \mathrm{sign}\Big(\mathrm{Tr}(W^{\mathrm{T}} X_i)\Big) = \mathrm{sign}\Big(\sum_m \mathrm{Tr}(W_f^{[m]} W_w^{\mathrm{T}} X_i^{[m]})\Big), \tag{1}$$

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

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

$$\begin{aligned} \min_{W_w, W_f^{[v]}, W_f^{[z]}} & \phi(W_f^{[v]}, W_f^{[z]}) + \lambda \cdot r(W_w, W_f^{[v]}, W_f^{[z]}) \\ & + C \cdot l(W_w, W_f^{[v]}, W_f^{[z]}), \end{aligned} \tag{2}$$

where $\phi(\cdot)$ is a regularizer term for reducing noise in the projected data, $r(\cdot)$ is an additional regularizer term related

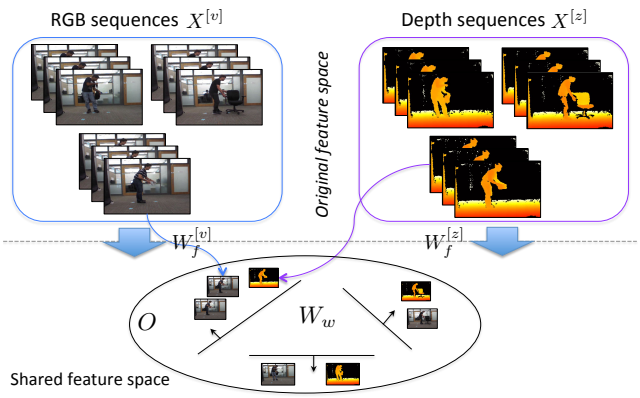


Figure 3. Graphical illustration of the proposed BHIM model. Parameter matrix $W_f^{[m]}$ $(m = 1, \cdots, M)$ projects the m modality data, $X^{[m]}$, into a learned shared space, and W_w is applied to classify the projected data regardless of modalities.

to the margin of the bilinear model, and $l(\cdot)$ computes the training loss for the two-modality data. λ and C are trade-off parameters balancing the importance of the corresponding terms.

Regularizer $\phi(W_f^{[v]}, W_f^{[z]})$ 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 knowledge between the two modalities. We define this term as

$$\phi(W_f^{[v]}, W_f^{[z]}) = I(X^{[v]}, O) + I(X^{[z]}, O), \qquad (3)$$

where $X^{[m]} = \{X_i^{[m]}\}_{i=1}^N$ $(m = v$ or $m = z)$ represents a set of all training samples in the m modality, $O = \frac{1}{2}(X^{[v]}W_f^{[v]} + X^{[z]}W_f^{[z]}) \in \mathcal{O}$ is the learned low-dimensional cross-modal features in the shared space, and $I(\cdot,\cdot)$ computes mutual information.

Cross-modal knowledge can be introduced to the model through the learning of the intermediate features O. Cross-modal features O inherit information from both RGB and depth data, including motion, 3D structural, and spatiotemporal relationship information. 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.

In addition, the term $\phi(W_f^{[v]}, W_f^{[z]})$ helps to reduce noise and produces a compact representation for cross-modal features O. In the learning of cross-modal features O, a large amount of noise irrelevant to action labels would also be introduced to the shared space, and thus degrades the recognition performance. By minimizing $\phi(W_f^{[v]}, W_f^{[z]})$, both noisy and discriminative information in O will be reduced, but discriminative information can be well captured by regularizer $r(W_w, W_f^{[v]}, W_f^{[z]})$ in Eq. (4). Parameter λ for regularizer $r(W_w, W_f^{[v]}, W_f^{[z]})$ is used for balancing the importance of the noise filter in BHIM.

Regularizer $r(W_w, W_f^{[v]}, W_f^{[z]})$ is used to measure the margin of the bilinear classifier. Minimizing $r(W_w, W_f^{[v]}, W_f^{[z]})$ is equivalent to maximizing the margin of the bilinear model, thereby capturing discriminative information. We define this term as

$$\begin{aligned} r(W_w, W_f^{[v]}, W_f^{[z]}) = &\frac{1}{2}\operatorname{Tr}(W_w W_f^{[v]\mathrm{T}} W_f^{[v]} W_w^{\mathrm{T}}) \\ &+ \frac{1}{2}\operatorname{Tr}(W_w W_f^{[z]\mathrm{T}} W_f^{[z]} W_w^{\mathrm{T}}). \end{aligned} \qquad (4)$$

Regularizer term $r(W_w, W_f^{[v]}, W_f^{[z]})$ naturally induces a low-rank classifier with the maximum rank of d. This restricts the degree of freedom of model parameter matrices. As shown in [22], the VC-dimension of low-rank classification models is proved to be less than that of the concatenated linear models.

Regularizer $r(W_w, W_f^{[v]}, W_f^{[z]})$ is minimized to extract discriminative information from cross-modal features O for action recognition. It works together with $\phi(W_f^{[v]}, W_f^{[z]})$ in Eq. (3) to extract discriminative information and filter out noise for recognition.

Loss function $l(W_w, W_f^{[v]}, W_f^{[z]})$ computes training loss given the learned model parameter matrices. We consider a binary classifier in this work, and define a hinge loss function for each modality, which is similar to the one in the binary SVM:

$$\begin{aligned} l(W_w, W_f^{[v]}, W_f^{[z]}) = \sum_i \Big[&\max(0, 1 - y_i \operatorname{Tr}(W_f^{[v]} W_w^{\mathrm{T}} X_i^{[v]}) \\ &+ \max(0, 1 - y_i \operatorname{Tr}(W_f^{[z]} W_w^{\mathrm{T}} X_i^{[z]}) \Big]. \end{aligned} \qquad (5)$$

Plugging Eq. (3), Eq. (4), and Eq. (5) into Eq. (2), optimal parameter matrices $W_f^{[v]}$, $W_f^{[z]}$ and W_w can be learned by the following constrained optimization problem:

$$\begin{aligned} \min_{W_w, W_f^{[v]}, W_f^{[z]}} \; &\sum_m \Big[I(X^{[m]}, O) + \frac{1}{2}\lambda \cdot \operatorname{Tr}(W_w W_f^{[m]\mathrm{T}} W_f^{[m]} W_w^{\mathrm{T}}) \\ &+ C \cdot \sum_i \xi_i^{[m]} \Big], \\ \text{s.t.} \quad &y_i \operatorname{Tr}(W_f^{[m]} W_w^{\mathrm{T}} X_i^{[m]}) \geqslant 1 - \xi_i^{[m]}, \quad \forall i, \forall m, \\ &\xi_i^{[m]} \geqslant 0, \quad \forall i, \forall m, \end{aligned} \qquad (6)$$

where $\xi_i^{[m]}$ is a slack variable for the m modality in the i-th RGB-D video.

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 method[1] [19, 4]. The bundle method is adopted as the inner problem solver due to its efficiency and good convergence.

We first reformulate the optimization problem (6) as an unconstrained regularized risk minimization problem:

$$\min_{W_w, W_f^{[v]}, , W_f^{[z]}} C \cdot \sum_i \sum_m L_i^{[m]} + \sum_m R^{[m]}, \qquad (7)$$

where

$$\begin{aligned} L_i^{[m]} &= \max(0, 1 - y_i \mathrm{Tr}(W_f^{[m]} W_w^{\mathrm{T}} X_i^{[m]})), \\ R^{[m]} &= I(X^{[m]}, O) + \frac{1}{2}\lambda \cdot \mathrm{Tr}(W_w W_f^{[m]\mathrm{T}} W_f^{[m]} W_w^{\mathrm{T}}), \end{aligned} \qquad (8)$$

are empirical loss and regularizer, respectively.

We solve the above problem by a coordinate descent algorithm. Specifically, if $W_f^{[m]}$ is fixed, the optimization problem is

$$\begin{aligned} &\min_{W_w} \frac{1}{2}\lambda \sum_m \mathrm{Tr}(W_w W_f^{[m]\mathrm{T}} W_f^{[m]} W_w^{\mathrm{T}}) \\ &+ C \sum_i \sum_m \max(0, 1 - y_i \mathrm{Tr}(W_f^{[m]} W_w^{\mathrm{T}} X_i^{[m]})). \end{aligned} \qquad (9)$$

To efficiently solve this problem, we define $A = \sum_m W_f^{[m]\mathrm{T}} W_f^{[m]}$, and define two auxiliary variables $\widehat{W}_w = W_w A^{\frac{1}{2}}$ and $\widehat{X}_i^{[m]} = X_i W_f^{[m]} A^{-\frac{1}{2}}$. Note that A is a matrix of size $d \times d$ that is in general invertible for small d. Then the problem (9) can be equivalently rewritten as

$$\min_{\widehat{W}_w} \frac{1}{2}\lambda \mathrm{Tr}(\widehat{W}_w^{\mathrm{T}} \widehat{W}_w) + C \sum_i \sum_m \max(0, 1 - y_i \mathrm{Tr}(\widehat{W}_w^{\mathrm{T}} \widehat{X}_i^{[m]})). \qquad (10)$$

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 $W_w = \widehat{W}_w A^{-\frac{1}{2}}$.

When W_w is fixed, $W_f^{[m]}$ for each modality can be optimized in a similar form to Eq. (7) and (8) but with W_w as constant. We define $B = W_w^{\mathrm{T}} W_w$, and further define two auxiliary variables, $\widetilde{W}_f$ and $\widetilde{X}_i$, as $\widetilde{W}_f^{[m]} = W_f^{[m]} B^{\frac{1}{2}}$ and $\widetilde{X}_i^{[m]} = X_i^{[m]\mathrm{T}} W_w B^{-\frac{1}{2}}$. Then, the parameter matrix $\widetilde{W}_f^{[m]}$ for each modality can be optimized independently by

$$\begin{aligned} &\min_{\widetilde{W}_f^{[m]}} \frac{1}{2}\mathrm{Tr}(\widetilde{W}_f^{[m]\mathrm{T}} \widetilde{W}_f^{[m]}) + \lambda I(\widetilde{X}^{[m]}, \widetilde{O}) \\ &\quad + C \sum_i \max(0, 1 - y_i \mathrm{Tr}(\widetilde{W}_f^{[m]\mathrm{T}} \widetilde{X}_i^{[m]})), \end{aligned} \qquad (11)$$

[1] https://forge.lip6.fr/projects/nrbm

Algorithm 1 Bilinear IB model learning algorithm

Input: $\{(X_i^{[m]}, y_i)\}_{i=1}^N (m = 1, \ldots, M)$.
Output: $W_w, W_f^{[m]}$.
Initialize variables $W_w, W_f^{[m]}$.
repeat
  Compute $A = \sum_m W_f^{[m]\mathrm{T}} W_f^{[m]}$, $\widehat{W}_w = W_w A^{\frac{1}{2}}$, and $\widehat{X}_i^{[m]} = X_i W_f^{[m]} A^{-\frac{1}{2}}$.
  Fix $W_f^{[m]}$, and optimize $\widehat{W}_w$ by (10).
  Recover $W_w = \widehat{W}_w A^{-\frac{1}{2}}$.
  Compute $B = W_w^{\mathrm{T}} W_w$, $\widetilde{W}_f^{[m]} = W_f^{[m]} B^{\frac{1}{2}}$, and $\widetilde{X}_i^{[m]} = X_i^{[m]\mathrm{T}} W_w B^{\frac{1}{2}}$.
  Fix W_w, and optimize $\widetilde{W}_f^{[v]}$ and $\widetilde{W}_f^{[z]}$ independently by (11).
  Recover $W_f^{[v]} = \widetilde{W}_f^{[v]} B^{-\frac{1}{2}}$ and $W_f^{[z]} = \widetilde{W}_f^{[z]} B^{-\frac{1}{2}}$.
until Objective changes $<$ threshold.

with the assumption that the conditional distribution $p(W_w, B^{-\frac{1}{2}} | X^{[m]}, O)$ is a uniform distribution[2]. This is also an unconstrained regularized risk minimization problem for linear SVM and can be solved by a bundle algorithm if $\widetilde{W}_f^{[m]}$ and $\widetilde{X}_i^{[m]}$ 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 $\widetilde{W}_f^{[m]}$, $W_f^{[m]}$ can be recovered by $W_f^{[m]} = \widetilde{W}_f^{[m]} B^{-\frac{1}{2}}$.

The proposed BHIM is solved by iteratively optimizing problems (10) and (11) until convergence. This is a biconvex problem as optimizing one parameter matrix holding the others fixed is a convex problem. The learning algorithm is shown in Algorithm 1.

3.3. Discussion

We highlight key properties of the proposed BHIM here.

Matrix form feature representation. Visual and depth features are represented in a matrix form in BHIM, which naturally considers spatiotemporal motion relationships of body parts. However, the relationships would be collapsed in a vector form representation in existing methods [13, 25].

Low-rank bilinear model. BHIM naturally models feature matrices using two model parameter matrices W_f and W_w. The rank of the proposed model is minimized to provide a better generalization power [22].

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 BHIM learns cross-modal features from RGB and depth data. The cross-modal fea-

[2] Please refer to supplemental material for details.

Table 1. Comparison results with various dimensionality d of the feature space. The dimensionality of features for each modality in linear SVM is $n_{xyt} \cdot d$.

Methods	$d = 1$			$d = 5$			$d = 31$		
	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	53.89%	50.00%	70.56%	90.00%	87.22%	91.11%	92.78%	80.00%	96.11%
Our method	83.33%	91.11%	96.11%	88.33%	76.11%	98.33%	93.89%	97.22%	100%

tures 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 $W_f^{[v]}$ or $W_f^{[z]}$ transfers information from original data $X^{[m]}$ to the learned shared features O. This helps exploit cross-modal knowledge if one modality is missing in testing.

4. Experiments

4.1. Datasets and Settings

The proposed method is evaluated on the MSR Action Pairs dataset [13] and MSR Daily Activity dataset [21]. 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 3 times, to provide a total of 360 videos for each of the RGB and depth modality. MSR Daily Activity dataset contains 16 types of activities 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. MSR Action Pairs Dataset

Videos in this dataset are temporally normalized to 10 frames with spatial resolution of 120×160. Histograms of gradient oriented feature is extracted from both depth and RGB videos with patch size 8×8. Thus, a total of $n_{xyt} = 3000$ patches are extracted from each video, with the feature dimensionality of $n_f = 31$. We follow [13] and use RGB-D videos of the first 5 subjects as training data.

Comparison experiment. We compare with existing method [13, 21, 26, 25, 7], and use linear SVM as baseline. We also extend the bilinear SVM [14] to capture two-modality data, and use it as baseline.

Results in Table 2 show that our method outperforms all the comparison approaches. We achieve 100% accuracy as we effectively use both visual and depth features. Compared with linear SVM that simply concatenates the two features into a long vector, our method finds the optimal space for fusing the two features, and thus improves the performance. Although bilinear SVM also learns a shared feature space for the two features, our method uses the information measure $\phi(W_f^{[v]}, W_f^{[z]})$ in Eq. (3) to compress data and reduce noise irrelevant to our recognition task. Our method also outperforms [13, 21, 26, 25], 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 learned features and parameter matrices are visualized in Figure 4.

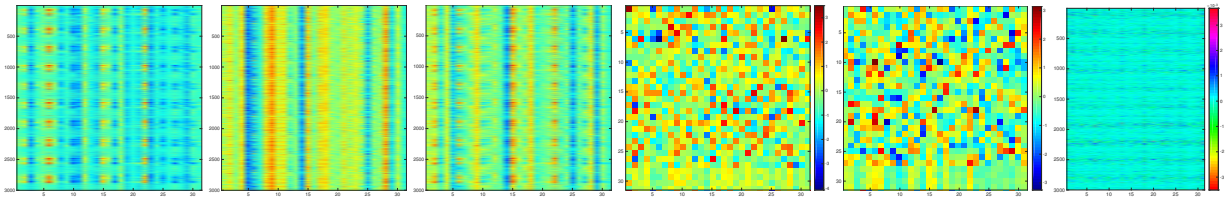
Table 2. Recognition accuracy of comparison methods on MSR Action Pairs dataset.

Methods	Accuracy
linear SVM	87.78%
Bilinear SVM [14]	96.11%
Deep Motion Maps [26]	66.11%
Skeleton+LOP+Pyramid [21]	82.22%
LTTL [7]	91.48%
HON4D [13]	96.67%
SNV [25]	98.89%
Our method	**100%**

Sensitivity to parameters. The proposed BHIM has three parameters to set, the maximum rank of the bilinear model d, the parameter C and the parameter λ in Eq. (2). In this experiment, we investigate the sensitivity of BHIM to these parameters.

We first test the sensitivity of BHIM to the maximum rank d. BHIM is compared with linear SVM and bilinear SVM with various d values. Note that there are a total of $n_{xyt} \times d$ elements in the shared space for each modality in BHIM 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 $n_{xyt} \cdot d$, making sure all the three 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 BHIM, the original feature matrix $X^{[m]}$ is projected by $W_f^{[m]}$. The rank parameter d is set to 1, 5, and 31, respectively.

The performance of the three methods on depth features, RGB features, and RGB-D features are shown in Table 1. Results indicate that our method achieves higher performance in most of the cases given low-dimensional features,



(a) Projected RGB features $X^{[v]}W_f^{[v]}$. (b) Projected depth features $X^{[z]}W_f^{[z]}$. (c) Learned shared features O. (d) Parameter matrix $W_f^{[v]}$. (e) Parameter matrix $W_f^{[z]}$. (f) Parameter matrix W_w.

Figure 4. Visualizations of (a) the projected visual features $X^{[v]}W_f^{[v]}$, (b) the projected depth features $X^{[z]}W_f^{[z]}$, (c) the learned cross-modal features O in the shared space, and the parameter matrices (d) $W_f^{[v]}$, (e) $W_f^{[z]}$, and (f) W_w.

Table 3. Knowledge transfer results on MSR Action Pairs dataset. $X \rightarrow Y$ denotes that X is the training data and Y is the testing data. $d = 31$ for both bilinear SVM and BHIM, and dimensionality of features in linear SVM is $n_{xyt} \cdot d$. The number of elements in the input feature vector/matrix to the three methods is the same.

Methods	RGB-D→RGB	RGB-D→Depth	RGB→RGB-D	Depth→RGB-D
linear SVM	83.33%	81.67%	87.22%	86.11%
Bilinear SVM	90.56%	93.89%	81.67%	91.67%
Our method	97.78%	92.78%	97.78%	93.33%

and its performance on RGB-D data is not sensitive to parameter d. When $d = 1$, the projected feature matrices may lose certain amount of information. However, the structural information is reserved in BHIM, resulting in significant higher performance over linear SVM. In addition, the learned shared space in BHIM is optimized for classification, while it is not the case in PCA. Compared with bilinear SVM, noisy information is reduced in BHIM, and thus it achieves superior performance.

When $d = 31$, even though linear SVM captures full information from visual and depth features, it does not capture spatiotemporal relationship information due to its vector form feature representation. In addition, depth and RGB features are concatenated in linear SVM, 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 BHIM solves these two problems by a matrix form feature representation and learning a shared feature space. The matrix form representation naturally captures spatiotemporal body part correlations. The learning of a shared feature space allows us to effectively use the two types of features for recognition.

BHIM achieves lower results on depth-only and RGB-only data compared with bilinear SVM when $d = 5$. This is because the learned cross-modal features in BHIM loses too much discriminative information using the information measure $\phi(W_f^{[v]}, W_f^{[z]})$ in Eq. (3). However, when use both of the two modalities, BHIM outperforms bilinear SVM since the discriminative information missing in one modality can be complemented from the other available modality.

RGB-D action recognition results of BHIM with different values of parameter C are shown in Table 4. Results indicate that our BHIM is insensitive to parameter C when the value of parameter C is lower than 1. However, the performance drops when the value becomes large.

Table 4. RGB-D action recognition results of our BHIM on MSR Action Pairs dataset with different values of parameter C.

C value	$C = 0.01$	$C = 0.1$	$C = 1$	$C = 5$
Accuracy	97.22%	98.33%	97.22%	64.44%

We also evaluate the performance of BHIM given different values of λ. The value of λ is set to 0.01, 0.1, 1, and 10. Results in Table 5 indicate that BHIM is insensitive to parameter λ. The largest performance difference is about 5% between $\lambda = 0.01$ and $\lambda = 0.1$. The insensitivity of BHIM to parameter λ significantly saves time in parameter tuning.

Table 5. RGB-D action recognition results of our BHIM on MSR Action Pairs dataset with different values of parameter λ.

λ value	$\lambda = 0.01$	$\lambda = 0.1$	$\lambda = 1$	$\lambda = 10$
Accuracy	97.22%	98.33%	92.78%	95.00%

Knowledge Transfer. We evaluate the performance of our BHIM, and investigate the effectiveness of the cross-modal features and the information measure when one modality is missing in training or testing. BHIM

is tested in four scenarios: depth data are missing in testing (RGB-D→RGB), RGB data are missing in testing (RGB-D→Depth), depth data are missing in training (RGB→RGB-D), and RGB data are missing in training (Depth→RGB-D). We compare BHIM with linear SVM and bilinear SVM, and investigate how the knowledge transferred from observed modality influences the performance of the three methods.

Recognition results in Table 3 show that BHIM significantly outperforms linear and bilinear SVM in this knowledge transfer experiment. Our BHIM achieves significantly higher accuracy than linear SVM. This demonstrates the superiority of using a matrix form feature representation and the learned cross-modal features in BHIM. Compared with bilinear SVM, BHIM also achieves superior results in most cases. Thanks to the information measure in learning the projection matrices, BHIM is capable of reducing noise in learning the shared feature space, and thus outperforms bilinear SVM.

4.3. 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 [21]. BHIM is first compared with existing approaches [26, 11, 27, 13, 21, 25] on this dataset, and then evaluated given RGB, depth, and RGB-D data, respectively. Linear SVM and bilinear SVM are used as baseline.

Table 6. Recognition accuracy of comparison methods on MSR Daily Activity Dataset.

Methods	Accuracy
linear SVM	65.00%
Bilinear SVM	85.63%
Depth Motion Maps [26]	43.13%
RGGP [11]	72.10%
Moving Pose [27]	73.80%
Local HON4D [13]	80.00%
Actionlet Ensemble [21]	85.75%
SNV [25]	86.25%
Our method	**86.88%**

BHIM is compared with existing approaches [26, 11, 27, 13, 21, 25], and results are shown in Table 6. BHIM achieves superior performance over state-of-the-art methods. BHIM significantly outperforms linear SVM possibly due to the learning of a shared feature space for the two types of features, and a matrix form representation that naturally captures spatiotemporal structural information. Recognition accuracy of BHIM is also higher than bilinear SVM due to the use of information measure, which is helpful in removing redundant information and noise. BHIM outperforms recent surface normal-based approaches [13, 25]. 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. BHIM achieves better performance than the actionlet ensemble approach [21] since we elegantly use visual and depth information, and effectively compress informative cues and remove noise before classification.

Performance of the proposed BHIM on the RGB-only, depth-only, and RGB-D data in the MSR Daily Activity dataset is also reported in this paper. Linear SVM and bilinear SVM are adopted as baseline. Recognition accuracy in Table 7 shows that BHIM achieves satisfactory results even though only one modality of features is given. When only depth features are given, linear SVM simply uses the features in the original feature space for classification. By contrast, our BHIM finds a better feature space to remove noise in order to achieve better performance. Compared with bilinear SVM, BHIM also utilizes information measure to compress data, and elegantly reduces redundancy in the data, which facilitates the recognition task.

Table 7. Comparison results on MSR Daily Activity Dataset given depth-only, RGB-only, and RGB-D data.

Methods	Depth	RGB	Depth+RGB
linear SVM	61.88%	54.38%	65.00%
Bilinear SVM	72.50%	67.50%	81.88%
Our method	81.88%	77.50%	86.88%

5. Conclusion

We have proposed a bilinear heterogeneous information machine (BHIM) for action recognition from RGB-D sequences. Both RGB and depth data are effectively utilized, and used to learn cross-modal features for recognition. We represent both visual and depth features in a matrix form to capture spatiotemporal relationships. A novel low-rank bilinear classifier is proposed to naturally model these feature matrices. BHIM learns a shared space for fusing RGB and depth data, and produces the cross-modal features. A large amount of noise is reduced in BHIM using the information measure. Classification is performed in the shared space using the learned cross-modal features. We learn a low-rank BHIM 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 BHIM is extensively evaluated on two public RGB-D action datasets, and outperforms state-of-the-art approaches.

Acknowledgement

This research is supported in part by the 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

[1] A. Argyriou, T. Evgeniou, and M. Pontil. Convex multi-task feature learning. *IJCV*, 2008.

[2] L. Bo, K. Lai, X. Ren, and D. Fox. Object recognition with hierarchical kernel descriptors. In *CVPR*, June 2011.

[3] L. Chen, W. Li, and D. Xu. Recognizing RGB images by learning from RGB-D data. In *CVPR*, 2014.

[4] T.-M.-T. Do and T. Artieres. Large margin training for hidden markov models with partially observed states. In *ICML*, 2009.

[5] S. Hadfield and R. Bowden. Hollywood 3D: Recognizing actions in 3D natural scenes. In *CVPR*, 2013.

[6] S. Ji, W. Xu, M. Yang, and K. Yu. 3D convolutional neural networks for human action recognition. *PAMI*, 2013.

[7] C. Jia, Y. Kong, Z. Ding, and Y. Fu. Latent tensor transfer learning for RGB-D action recognition. In *ACM Multimedia*, 2014.

[8] T. Kobayashi. Low-rank biliner classification: Efficient convex optimization and extensions. *IJCV*, 2014.

[9] Y. Kong, Y. Jia, and Y. Fu. Interactive phrases: Semantic descriptions for human interaction recognition. In *PAMI*, 2014.

[10] W. Li, Z. Zhang, and Z. Liu. Action recognition based on a bag of 3D points. In *CVPR workshop*, 2010.

[11] L. Liu and L. Shao. Learning discriminative representations from RGB-D video data. In *IJCAI*, 2013.

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

[13] O. Oreifej and Z. Liu. HON4D: Histogram of oriented 4D normals for activity recognition from depth sequences. In *CVPR*, 2013.

[14] H. Pirsiavash, D. Ramanan, and C. Fowlkes. Bilinear classifiers for visual recognition. In *NIPS*, 2009.

[15] M. Raptis and L. Sigal. Poselet key-framing: A model for human activity recognition. In *CVPR*, 2013.

[16] J. Shotton, R. Girshick, A. Fitzgibbon, T. Sharp, M. Cook, M. Finocchio, R. Moore, P. Kohli, A. Criminisi, A. Kipman, and A. Blake. Efficient human pose estimation from single depth images. *PAMI*, 2013.

[17] K. Tang, L. Fei-Fei, and D. Koller. Learning latent temporal structure for complex event detection. In *CVPR*, 2012.

[18] J. B. Tenenbaum and W. T. Freeman. Separating style and content with bilinear models. *Neural Computation*, 2000.

[19] C. H. Teo, Q. Le, A. Smola, and S. Vishwanathan. A scalable modular convex solver for regularized risk minimization. In *KDD*, 2007.

[20] N. Tishby, F. C. Pereira, and W. Bialek. The information bottleneck method. In *Proc. of the 37-th Annual Allerton Conference on Communication, Control and Computing*, pages 368–377, 1999.

[21] J. Wang, Z. Liu, Y. Wu, and J. Yuan. Mining actionlet ensemble for action recognition with depth cameras. In *CVPR*, June 2012.

[22] L. Wolf, H. Jhuang, and T. Hazan. Modeling appearances with low-rank svm. In *CVPR*, 2007.

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

[24] C. Xu, D. Tao, and C. Xu. Large-margin multi-view information bottleneck. *PAMI*, 36(8), 2014.

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

[26] X. Yang, C. Zhang, and Y. Tian. Recognizing actions using depth motion maps-based histograms of oriented gradients. In *ACM Multimedia*, 2012.

[27] M. Zanfir, M. Leordeanu, and C. Sminchisescu. The moving pose: An efficient 3D kinematics descriptor for low-latency action recognition and detection. In *ICCV*, 2013.