Depth and Skeleton Associated Action Recognition without Online Accessible RGB-D Cameras

Yen-Yu Lin[1] Ju-Hsuan Hua[2*] Nick C. Tang[1] Min-Hung Chen[1] Hong-Yuan Mark Liao[1]
[1]Academia Sinica, Taiwan [2]Carnegie Mellon University, USA

Abstract

The recent advances in RGB-D cameras have allowed us to better solve increasingly complex computer vision tasks. However, modern RGB-D cameras are still restricted by the short effective distances. The limitation may make RGB-D cameras not online accessible in practice, and degrade their applicability. We propose an alternative scenario to address this problem, and illustrate it with the application to action recognition. We use Kinect to offline collect an auxiliary, multi-modal database, in which not only the RGB videos but also the depth maps and skeleton structures of actions of interest are available. Our approach aims to enhance action recognition in RGB videos by leveraging the extra database. Specifically, it optimizes a feature transformation, by which the actions to be recognized can be concisely reconstructed by entries in the auxiliary database. In this way, the inter-database variations are adapted. More importantly, each action can be augmented with additional depth and skeleton images retrieved from the auxiliary database. The proposed approach has been evaluated on three benchmarks of action recognition. The promising results manifest that the augmented depth and skeleton features can lead to remarkable boost in recognition accuracy.

1. Introduction

Computer vision techniques are highly adapted to available imaging devices. We are aware of the recent advances in imaging devices, such as the RGB-D camera *Microsoft Kinect*, the binocular camera *FUJIFILM FinePix Real 3D*, and the lightfield camera *Lytro*. The images they record provide rich and diverse information. Thus, these emerging cameras complement the conventional 2D RGB cameras in developing computer vision algorithms. A vast amount of research effort has been made on investigating the multi-modal images taken by both conventional and emerging cameras, with the aim to better solve computer vision tasks and even initiate new applications.

Action recognition is one of the computer vision applications that can be facilitated by emerging imaging devices, especially the RGB-D cameras. As pointed out in [21], the *large intra-class variations*, such as different poses and partial occlusions, make automatic action recognition very challenging. Information carried by RGB images is in general insufficient to account for the unfavorable variations. Recent research efforts, *e.g.*, [9, 24, 27], have demonstrated that the depth images taken by Kinect as well as the inferred human 3D skeleton structures are very useful for handling the intra-class variations and building up a more accurate system of action recognition. However, RGB-D cameras usually have short effective distances. For instance, Kinect is with the effective range between $1.2 \sim 3.5$ meters. The limitation makes Kinect not online accessible in many real-world applications, such as surveillance.

We address this issue, and consider an alternative scenario, in which an auxiliary, multi-modal database is established by Kinect in advance. The database contains actions of interest, and its entries are in form of triplets: the RGB videos, the depth maps, and the skeleton structures of human actions. Our focus now is how to accurately recognizing actions in RGB videos by leveraging this extra database. Figure 1a shows the problem setting of this work.

We treat actions to be recognized, called target actions throughout, as queries to the auxiliary database. Our goal is to retrieve the appropriate depth maps and skeleton structures for them. Before that, we should account for the variations between the target actions and the actions in the auxiliary database. The proposed approach deals with the two tasks, *cross-modal retrieval* and *domain adaptation*, at the same time. By assuming that the inter-database variations can be modeled by a linear transformation, our approach optimizes a feature transformation, by which the transformed target actions can be well reconstructed by the actions in the auxiliary database. We cast this task as a rank minimization problem based on the formulation of Jhuo *et al*. [13], with the extension to carry out discriminant reconstruction. The resulting constrained optimization problem can be effectively solved by *augmented Lagrange multiplier* (ALM) method [15].

*This work was performed while this author was at Academia Sinica.

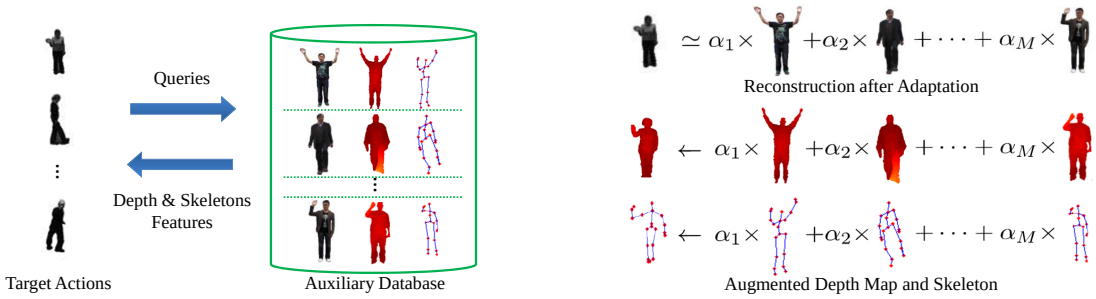


Figure 1. (a) We aim to improve action recognition in RGB videos by leveraging an auxiliary, multi-modal database. (b) Our approach deals with domain adaptation and cross-modal feature retrieval at the same time. Each action to be recognized is augmented with the additional depth and skeleton features that are retrieved from the auxiliary database.

After completing the optimization, domain adaptation is carried out. More importantly, each target action can be augmented with additional depth and skeleton features by sharing the optimized reconstruction coefficients. Namely, how the depth and skeleton features are retrieved for a target action is the same as how its RGB features are reconstructed. This idea is shown in Figure 1b. Then, *multiple kernel learning* (MKL) is adopted to fuse the original RGB features and the augmented depth and skeleton features, and to achieve better performance.

The main contribution of this work is to provide an alternative way of utilizing RGB-D cameras to facilitate action recognition even when they are not online accessible. The proposed approach is comprehensively evaluated on three benchmarks of action recognition, which were taken in diverse environments and contains actions of different classes. By using the same auxiliary database, our approach leads to remarkable accuracy improvement on each benchmark. Besides, our approach is designed in a general manner, and can be applied to vision applications where multi-modal images are helpful.

2. Related Work

Human action recognition has received strong attention in computer vision. Being one of the most important components in computer vision, action recognition is essential to widespread applications, such as surveillance and human-computer interaction. As indicated in [21], one fundamental difficulty of action recognition results from the large intra-class variations. The variations are caused by both intrinsic and extrinsic factors, such as posture differences among subjects, cluttered backgrounds, different camera perspectives, mutual or self occlusions.

Designing powerful feature descriptors against intra-class variations has gained significant progress. *Global descriptors*, *e.g.*, [8], encode the whole action observations. Despite the compactness and simplicity, global descriptors are often very sensitive to occlusions and deformations. *Local descriptors*, *e.g.*, [14, 17], instead treat an action as a gather of patches or spatio-temporal cubes. Local descriptors are widely used in the representation of actions. However, the geometric structure among local features is ignored. It may result in performance degradation.

To address this issue, one research trend of action recognition focuses on modeling geometric relationships among local features. For instance, Matikainen *et al*. [18] generated a frequency lookup table to specify the geometrical displacements between local features. Prabhakar *et al*. [22] estimated the causalities between visual words, and included them as parts of the features. Besides, graphical models, such as *factorial conditional random fields* [28] and *hidden Markov model* [6], have been applied to formulating the spatio-temporal correlation of local evidences. The foregoing approaches to action recognition were developed upon RGB images/videos. Their performances are bounded by the available information in RGB images/videos.

Owing to the recent advances in sensor technology, it has been feasible to record color as well as depth images in real time by RGB-D cameras, *e.g.*, Kinect. Research efforts have shown that depth maps afford informative clues for human pose estimation [9, 24] and action recognition [19, 33]. Besides, the *OpenNI library*[1] was developed upon RGB-D cameras, and can identify the positions of key joints on the human body, *i.e.*, the skeleton. Researches, *e.g.*, [27], on 3D skeleton representation and correction open the opportunity of handling multi-view action recognition. The introduction of depth and skeleton information indeed benefits action recognition. However, the short ranges of effective distances still make RGB-D cameras inapplicable in mean real-world applications

Training data acquisition is also a challenge for action and object recognition. Sufficient training data are required to derive a good predictor. In practice, complete training dataset may not be available, and labeling huge data is also expensive. *Transfer learning* [20] can alleviate the afore-

[1] http://www.openni.org/

mentioned problems. Cao *et al*. [5] leveraged the information from a source dataset, and reduced the number of the training data required in model learning. Yin *et al*. [34] associated samples to be recognized with alike data from an extra dataset, and exploited the expanded data to make better predictions. Jhuo *et al*. [13] instead enriched the set of training data from multiple sources via domain adaptation and data reconstruction. Chen and Grauman [7] synthesized training data from an unlabeled pool of videos, and facilitated the learning of action categories with static images. Our approach also utilizes knowledge transferred from an auxiliary database for improving action recognition, but it is different from the foregoing approaches in the sense that it borrows visual features across different video modalities, and resolves the problems caused by the absence of RGB-D cameras.

Multiple kernel learning (MKL) [1, 23] refers to learning a kernel machine with multiple kernels. MKL has shown its capability for fusing heterogenous features and leading to boosted performance. In our case, we represent actions described by each type of features as a kernel matrix. These features include the original RGB features as well as the augmented depth and skeleton features. MKL carries out discriminant learning, and derives an optimal kernel over a given convex set of kernels. It turns out that all the features are fused in the domain of kernel matrices.

3. Problem Definition

We aim to carry out depth- and skeleton-associated action recognition, even when there are no RGB-D cameras online available. The goal is realized by borrowing information from an auxiliary, multi-modal database in this work. Suppose that a set of actions is given, $D = \{\mathbf{x}_i, y_i\}_{i=1}^{N}$, where $\mathbf{x}_i \in \mathcal{X}$ and $y_i \in \mathcal{Y} = \{1, 2, ..., C\}$ are the RGB feature representation of the ith action and its class label. Besides, an auxiliary database, $\tilde{D} = \{(\tilde{\mathbf{x}}_i, \tilde{\mathbf{d}}_i, \tilde{\mathbf{s}}_i)\}_{i=1}^{M}$, taken by Kinect is available, where $\tilde{\mathbf{x}}_i \in \mathcal{X}$, $\tilde{\mathbf{d}}_i \in \mathcal{D}$, and $\tilde{\mathbf{s}}_i \in \mathcal{S}$ are the RGB, depth, and skeleton features of the ith instance, respectively. The data in $\tilde{D}$ are unlabeled. We use *tildes* to mark the auxiliary instances and dataset for clearness. D and $\tilde{D}$ have common RGB data domain $\mathcal{X}$, but have different data distributions. This is why domain adaptation is required. We focus on associating each action $\mathbf{x}_i \in D$ with appropriate depth map $\mathbf{d}_i$ and skeleton structure $\mathbf{s}_i$, so that the absence of RGB-D cameras is compensated.

The auxiliary dataset $\tilde{D}$ is compiled to cover the actions of interest in advance, *i.e*., $\mathcal{Y}$ in this case. Collecting $\tilde{D}$ offline is reasonable since we often focus on recognizing predefined types of actions in most applications. Nevertheless, the classes of actions in D are not required to be the same as those in $\tilde{D}$. In our experiments, D is in turn one of the three action recognition benchmarks, while $\tilde{D}$ we collect contains all types of actions included in the three benchmarks.

4. The Proposed Approach

The proposed approach is described in this section.

4.1. Domain adaptation

To borrow features from $\tilde{D}$ to D, we correlate the two independently collected datasets by exploring their common video modality, RGB. An intuitive way, like [34], is the nearest neighbor search in the RGB domain. Then, the corresponding depth and skeleton features are associated. However, this method ignores the inter-database variations, and results in suboptimal performance in our cases.

Domain adaptation is applied to tackling the variations. Among the existing algorithms of domain adaptation, we prefer those that work in an interpretable way, since we would like to reapply the obtained adaptation model to the other video modalities. Inspired by the good performance reported in [13], we also adopt reconstruction-based domain adaptation, so that the optimized reconstruction coefficients can be reused in the synthesis of the augmented depth and skeleton features. By assuming that the inter-database variations between D and $\tilde{D}$ can be modeled by a linear transformation, the transformed actions in D will be well reconstructed by actions in $\tilde{D}$. The idea can be specified by:

$$WX = \tilde{X}A + E, \tag{1}$$

where $X = [\mathbf{x}_1 \cdots \mathbf{x}_N] \in \mathbb{R}^{d \times N}$ and $\tilde{X} = [\tilde{\mathbf{x}}_1 \cdots \tilde{\mathbf{x}}_M] \in \mathbb{R}^{d \times M}$ are the data matrices of D and $\tilde{D}$, respectively, $W \in \mathbb{R}^{d \times d}$ is the linear transformation, $A = [\boldsymbol{\alpha}_1 \cdots \boldsymbol{\alpha}_N] \in \mathbb{R}^{M \times N}$ is the matrix of reconstruction coefficients, and $E = [\mathbf{e}_1 \cdots \mathbf{e}_N] \in \mathbb{R}^{d \times N}$ is the error matrix. It can be checked column by column in (1) that each transformed sample $W\mathbf{x}_i$ is well reconstructed by auxiliary data $\tilde{X}\boldsymbol{\alpha}_i$, if residue $\mathbf{e}_i$ is minimized.

Although formulation (1) accounts for the inter-database variations, W can be optimized in a better way. First, distinct from the setting of [13], D is labeled multi-class data in our cases. We can activate discriminant learning for deriving better W. Second, outliers may exist in D. We need to handle the outliers, since they may dominate the reconstruction errors in optimizing W.

Without loss of generality, we assume training data are sorted according to their labels, *i.e*., $X = [X^1 \cdots X^C]$, where X^c is the submatrix whose columns are data of class c. $A = [A^1 \cdots A^C]$ and $E = [E^1 \cdots E^C]$ are defined accordingly. We consider discriminant learning and outlier handling, and reformulate (1) as the following constrained optimization problem:

$$\begin{aligned} \min_{W,A,E} \quad & \sum_{c=1}^{C} \operatorname{rank}(A^c) + \lambda \|E\|_{2,1} \\ s.t. \quad & WX = \tilde{X}A + E \text{ and } WW^\top = I, \end{aligned} \tag{2}$$

Algorithm 1: The inexact ALM algorithm for solving constrained optimization problem (4)

Input : Target actions X, Auxiliary actions $\tilde{X}$, Parameter λ;

Initialize: $E = 0,\ W = I,\ A = (\tilde{X}^\top \tilde{X})^{-1}\tilde{X}^\top WX,\ U = 0,\ V = 0,\ \mu = 10^{-3}$;

while *not converged* **do**

1. Update F by $F^c = \arg\min_{F^c} \frac{1}{\mu}\|F^c\|_* + \frac{1}{2}\|F^c - (A^c + \frac{U^c}{\mu})\|_F^2$, for $c = 1, 2, ..., C$;
2. Update W by $W = (\tilde{X}A + E - \frac{V}{\mu})X^\top(XX^\top)^{-1}$;
3. $W \leftarrow \text{orthogonal}(W)$;
4. Update E by $E = \arg\min_E \frac{\lambda}{\mu}\|E\|_{2,1} + \frac{1}{2}\|E - (WX - \tilde{X}A + \frac{V}{\mu})\|_F^2$;
5. Update A by $A = (I + \tilde{X}^\top\tilde{X})^{-1}[\tilde{X}^\top(WX - E) + \frac{1}{\mu}(\tilde{X}^\top V - U) + F]$;
6. Update the Lagrange multipliers: $U = U + \mu(A - F)$, $V = V + \mu(WX - \tilde{X}A - E)$;
7. Update the penalty parameter μ by $\mu = 1.2\mu$;
8. Check convergence conditions: $A - F \longrightarrow 0$ and $WX - \tilde{X}A - E \longrightarrow 0$;

where $\|E\|_{2,1} = \sum_{i=1}^{N} \|\mathbf{e}_i\|_2$ is the $l_{2,1}$ norm of E, and λ is a positive tradeoff parameter. Constraint $WW^\top = I$ ensures that W is a basis transformation. Minimizing the rank of A^c enforces that the reconstruction coefficient vectors corresponding to data of class c are similar to each other. It follows that discriminant learning is activated. The use of the $l_{2,1}$ norm in error measure alleviates the overfitting problem caused by the outliers.

Rank minimization in general is known as an NP-hard problem, and there is no efficient algorithm to solve it. Hence, we consider the convex relaxation of (2):

$$\begin{aligned} \min_{W,A,E} \quad & \sum_{c=1}^{C} \|A^c\|_* + \lambda\|E\|_{2,1} \\ s.t. \quad & WX = \tilde{X}A + E \ \text{ and } \ WW^\top = I, \end{aligned} \tag{3}$$

where $\|A^c\|_*$ is the *nuclear norm* of A^c, *i.e.*, sum of the singular values. It serves as a convex approximation of $\text{rank}(A^c)$. In this work, we solve the constrained optimization problem (3) by using the *Augmented Lagrange Multiplier* (ALM) method [15]. The ALM method deals with a constrained optimization problem by solving a series of unconstrained ones. To begin with, we introduce an auxiliary variable $F = [F^1 \cdots F^C] \in \mathbb{R}^{M \times N}$, and convert (3) into an equivalent form

$$\begin{aligned} \min_{W,A,F,E} \quad & \sum_{c=1}^{C} \|F^c\|_* + \lambda\|E\|_{2,1} \\ s.t. \quad & WX = \tilde{X}A + E \ \text{ and } \quad A = F. \end{aligned} \tag{4}$$

The orthogonality constraint of W is temporarily ignored. Nevertheless, W is orthogonalized afterwards as in most orthogonality preserving methods, *e.g.*, [31]. We optimize (4) by the inexact ALM method, which minimizes the *augmented Lagrange function* of (4):

$$\begin{aligned} \min_{W,A,F,E,U,V} & \sum_{c=1}^{C} \|F^c\|_* + \alpha\|E\|_{2,1} + \langle U, A - F\rangle + \frac{\mu}{2}\|A - F\|_F^2 \\ & + \langle V, WX - \tilde{X}A - E\rangle + \frac{\mu}{2}\|WX - \tilde{X}A - E\|_F^2, \end{aligned} \tag{5}$$

where $\langle \cdot, \cdot \rangle$ denotes the inner product operator, μ is a positive penalty parameter, and U and V are the Lagrange multipliers. Refer to [15] for the details of the augmented Lagrange function.

Starting with a small value of penalty parameter μ, the inexact ALM method iteratively solves (5) with gradually increased μ. The procedure is repeated until all the constraints in (4) are satisfied. At each iteration, the strategy of *alternate optimization* is adopted for solving variables $\{W, A, F, E\}$. Namely, we optimize one of the four variables by fixing the rest, and then switch roles of the variables sequentially. For the optimization problem w.r.t. $\{F^c\}_{c=1}^C$, the *singular value shrinkage operator* $\mathcal{D}_\tau$ in [4] is used as the solver to optimize F^c with $\mathcal{D}_{\frac{1}{\mu}}(A^c + \frac{U^c}{\mu}) = \arg\min_{F^c} \frac{1}{\mu}\|F^c\|_* + \frac{1}{2}\|F^c - (A^c + \frac{U^c}{\mu})\|_F^2$. For the optimization problems regarding W and A, there exist closed-form solutions derived by setting the partial derivative to zero. That is, $W = (\tilde{X}A + E - \frac{V}{\mu})X^\top(XX^\top)^{-1}$ and $A = (I + \tilde{X}^\top\tilde{X})^{-1}[\tilde{X}^\top(WX - E) + \frac{1}{\mu}(\tilde{X}^\top V - U) + F]$. Besides, we apply QR-decomposition to orthogonalize the obtained W such that constraint $WW^\top = I$ holds. As for the optimization problem w.r.t. E, it can be updated using the analytical solution in [16]. We summarize the optimization procedure in Algorithm 1.

The optimization procedure converges with $30 \sim 70$ iterations in our experiments. The step of optimizing W takes the most time. The running time can be significantly reduced after using principal component analysis (PCA) to preprocess data. Each iteration is executed within five seconds in all of our experiment settings.

4.2. Feature augmentation

After completing the optimization procedure in Algorithm 1, the linear transformation W which adapts the variations between D and $\tilde{D}$ is obtained. The next step is to explore how the transformed target actions are reconstructed by actions in the auxiliary database. The optimized A records the reconstruction coefficients for training data, but how unseen testing data are reconstructed remains unknown. To address this issue, we transform each training and testing action $\mathbf{x}$ via the learned W, and seek its reconstruction coefficients by solving

$$\boldsymbol{\alpha} = \arg\min_{\boldsymbol{\alpha}} \|W\mathbf{x} - \tilde{X}\boldsymbol{\alpha}\|^2 + \gamma\|\boldsymbol{\alpha}\|^2, \tag{6}$$

where γ is a positive constant, controlling the trade-off between the fitness and the regularization term. There exists closed-form solution to (6), *i.e.*,

$$\boldsymbol{\alpha} = (\tilde{X}^\top \tilde{X} + \gamma I)^{-1} \tilde{X}^\top W\mathbf{x}. \tag{7}$$

Once $\boldsymbol{\alpha}$ is obtained, we are ready to retrieve the depth and skeleton features of $\mathbf{x}$. Recall that the data entries of $\tilde{D}$ are in form of triplets. We exploit the modality correspondences in $\tilde{D}$, and associate target action $\mathbf{x}$ with additional depth features $\mathbf{d}$ and skeleton features $\mathbf{s}$ by the rule: How the depth and skeleton features are retrieved is the same as how the RGB features are reconstructed. That is,

$$\mathbf{d} \leftarrow [\tilde{\mathbf{d}}_1 \cdots \tilde{\mathbf{d}}_M]\boldsymbol{\alpha} \quad \text{and} \quad \mathbf{s} \leftarrow [\tilde{\mathbf{s}}_1 \cdots \tilde{\mathbf{s}}_M]\boldsymbol{\alpha}. \tag{8}$$

The association procedure in (8) is repeated for each training sample. It follows that the augmented training dataset $D = \{(\mathbf{x}_i, \mathbf{d}_i, \mathbf{s}_i)\}_{i=1}^N$ is constructed. The procedure is also applicable to testing samples. Parameter λ in (4) and γ in (7) are critical to the performance. Their values are determined by cross validation in our implementation.

4.3. Prediction with augmented features

In the augmented dataset, three video modalities of each action are available at the same time. Multiple kernel learning can then be adopted for combining the three heterogeneous features to achieve better performance. Specifically, we compile an kernel matrix for actions in each modeality. Kernel matrix $K_{\mathbf{x}} = [k_{\mathbf{x}}(\mathbf{x}_i, \mathbf{x}_j)] \in \mathbb{R}^{N\times N}$ is constructed for actions in RGB videos with

$$k_{\mathbf{x}}(\mathbf{x}_i, \mathbf{x}_j) = \exp\left(\frac{-\|\mathbf{x}_i - \mathbf{x}_j\|^2}{\sigma_{\mathbf{x}}^2}\right), \tag{9}$$

where $\sigma_{\mathbf{x}}$ is the hyperparameter. Accordingly, we have kernels $K_{\mathbf{d}}$ and $K_{\mathbf{s}}$ for actions in depth and skeleton features, respectively. In this work, we adopt *SimpleMKL* [23], one of the state-of-the-art MKL packages, to learn SVM classifiers with the three kernels as input. It will automatically determine the optimal combination of the three types of features in the domain of kernel matrices.

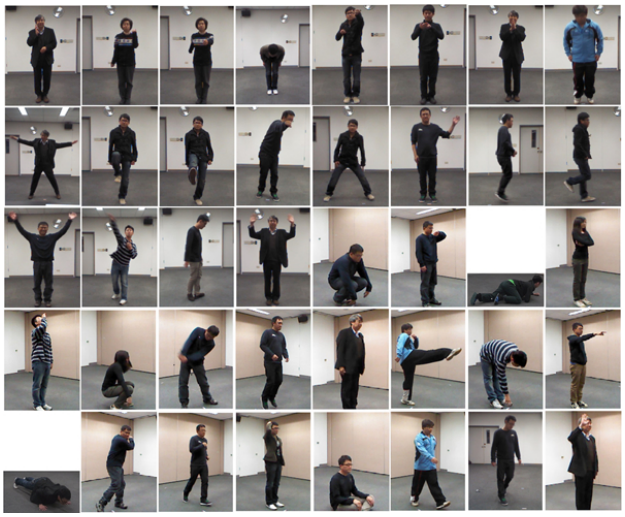


Figure 2. The auxiliary dataset we collected. One example comes from each of the 40 action categories.

5. Feature Representation

The features representations adopted to characterize actions in RGB videos, depth maps, and skeleton structures are described in this section.

RGB videos: The backgrounds of RGB videos are firstly estimated by using the inpainting technique [25]. We then take the background regions and adopt the background subtraction algorithm [2] to segment out the foreground region of each video frame. Based on the foregrounds, an RGB video is resized to $48\times64\times t$ pixels, where t is the number of frames. The *3D-HOG descriptor* [30] is applied to extract features in the space-time volume. In more detail, we use $16\times16\times16$ pixel blocks, each of which is further divided into $2\times2\times2$ cells. Five hundred prototypes are derived to build up the embedding space. It leads to a compact representation for RGB action videos.

Depth maps: We apply the *spatio-temporal local binary pattern* [35] to depict depth maps.

Skeleton structures: We implement *Fourier temporal pyramid* [27] to represent skeleton structures. The short time Fourier transform is applied to each skeleton segments in a 3-level pyramid. The feature representation is the concatenation of the Fourier coefficients from all the segments.

6. Experimental Results

In this section, the performance of our approach is evaluated on three action recognition benchmarks.

6.1. Datasets

Three benchmarks of action recognition, including IXMAS [29], i3DPost [10], and UIUC-1 [26], are adopted for performance evaluation. Besides, we use Microsoft Kinect

Method	Ours: d+s	Ours: d	Ours: s	RGB	Bor-DEP	Bor-SKE	KSDA	1NN-Bor	[32]
Accuracy	**89.1**	81.6	88.5	78.6	51.2	82.6	80.6	80.3	87.7

Table 1. Recognition rates (%) by different approaches on IXMAS dataset.

Method	Ours: d+s	Ours: d	Ours: s	RGB	Bor-DEP	Bor-SKE	KSDA	1NN-Bor	[12]
Accuracy	**88.3**	84.4	87.9	82.0	57.8	80.1	82.8	83.2	84.9

Table 2. Recognition rates (%) by different approaches on i3DPost dataset.

to build up a three-modal dataset, which serves as the auxiliary database in all the experiments on the three benchmarks. These databases are described as follows.

Auxiliary dataset: Ten actors were employed to construct the dataset. Each actor performed 40 types of actions to cover all the action classes in the three benchmarks used in the experiments. Each action was recorded by two cameras, respectively located with view angles of 0° and 45°. For each action, we further included its horizontal mirror in the dataset. Figure 2 gives an overview of the dataset.

IXMAS: We followed [32], in which 11 kinds of actions by 10 actors were used. We conducted performance evaluation and comparison on action videos captured by camera #1, #2, and #3 in the dataset.

i3DPost: For comparing with other state-of-the-art methods, we followed the evaluation protocols suggested in [12], where total 8 daily activities performed by 8 actors were considered. Our approach and the adopted baselines are evaluated on videos taken by camera #5 and #6.

UIUC-1: The UIUC-1 human activity dataset consists of 532 high resolution sequences of 14 activities performed by 8 actors. All the sequences are used in our experiments.

6.2. Baselines

We establish several baselines for performance comparison. The goal of comparison is to identify the contributions of the developed components in our approach. These baselines are denoted below in bold and in abbreviation:

RGB: This baseline simply ignores the information from the auxiliary database. It extracts the RGB features, and learns an SVM classifier to make the prediction. Comparing to this baseline examines whether the auxiliary database helps in recognizing target actions.

Bor-DEP: This baseline discards the original RGB features and simply works on the associated depth maps. Investigating its performance helps identify whether the associated features themselves are informative or not.

Bor-SKE: It is the same as Bor-DEP, except the used features become the associated skeleton structures.

KSDA: Kernel SDA [3] is a semi-supervised learning algorithm. In our cases, the RGB action videos in D and $\tilde{D}$ are considered as the labeled and the unlabeled training data in kernel SDA, respectively.

1NN-Bor: For each sample in D, we associate its nearest sample in $\tilde{D}$, and borrow the corresponding depth and skeleton features. MKL is used for feature fusion. This baseline neglects the possible inter-database variations.

Our approach is denoted by **Ours:d+s**, if it works on the RGB features and the associated depth and skeleton features. Two degenerate variants **Ours:d** and **Ours:s** are built when it considers the RGB features and either the associated depth or the associated skeleton features, respectively.

6.3. Experiment settings

To make a fair comparison, we adopt the widely used setup, *leave-one-actor-out* (LOAO) cross validation. Suppose that there are N actors in the benchmark. LOAO is the same as N-fold cross validation, except training data of the same actor must belong to an identical fold. The obtained recognition rates of our approach, the adopted baselines, and the state-of-the-art methods are reported in Table $1 \sim 3$, one for each benchmark.

We further test our approach with different numbers of actors used in training. Specifically, action videos by k actors are used for training, while the rest are for testing. We set $k = 3, ..., N-1$, respectively. In each case, we try N random splits of the actors, and compute the average recognition rates. Figure 3 shows the quantitative results on the three benchmarks.

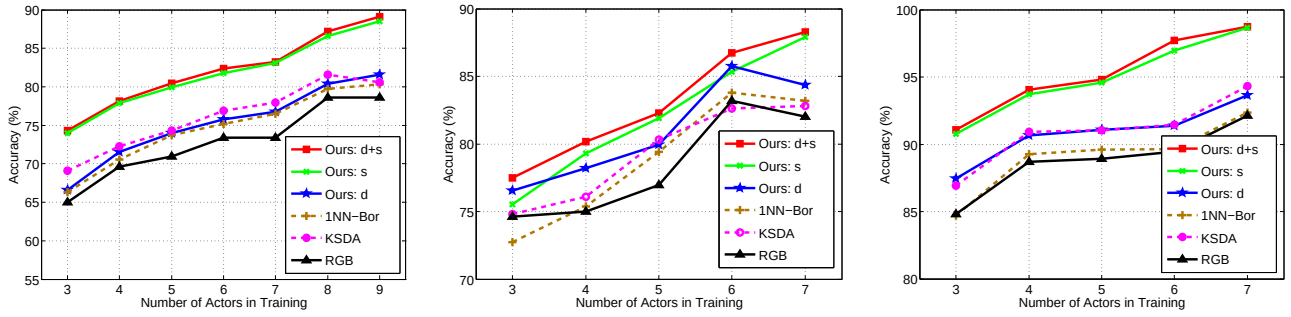
6.4. Analysis and discussion

Table 1 and Figure 3 give the results regarding IXMAS dataset. Baseline KSDA is a bit better than baseline RGB. It reveals that the unlabeled RGB videos in the auxiliary dataset provide useful information for the regularization of classifier training. The baselines Bor-DEP and Bor-SKE get recognition rates of 51.2% and 82.6% in the LOAO setup, respectively. As can be seen that the augmented skeleton features by the proposed mechanism are quite informative, and Bor-SKE is even superior to baseline RGB. Baseline 1NN-Bor ignores the inter-database variations, and results in suboptimal performance. Instead, our approach can make the most of the auxiliary database. It utilizes the augmented depth and skeleton features, and leads to a satisfactory recognition rate, 89.1%.

Similar observations can be found in the results on i3DPost and UIUC-1 datasets. It is still worth mentioning

Method	Ours: d+s	Ours: d	Ours: s	RGB	Bor-DEP	Bor-SKE	KSDA	1NN-Bor	[26]	[11]
Accuracy	98.7	93.6	98.7	92.1	74.2	95.0	94.3	92.4	98.3	**99.6**

Table 3. Recognition rates (%) by different approaches on UIUC-1 dataset.



(a) (b) (c)

Figure 3. The recognition rates of various approaches with different numbers of training actors on (a) IXMAS, (b) i3DPost, and (c) UIUC-1.

some interesting observations. First, the augmented depth features work better, and complement the original RGB features in i3DPost dataset. This may result from that self-occlusions frequently occur in i3DPost dataset, and depth maps helps in this situation. Second, it can be observed in Figure 3 that the introduction of the auxiliary database indeed compensates for the lack of training data. For instance, our approach learned with 3 training actors in IXMAS is still superior to baseline RGB with 7 training actors. Similar phenomena can be found in the other two datasets.

The performance gains of our approach over baseline RGB are very significant. In the LOAO setup, it is 10.5% $(= 89.1\% - 78.6\%)$ in IXMAS, 6.3% $(= 88.3\% - 82.0\%)$ in i3DPost, and 6.6% $(= 98.7\% - 92.1\%)$ in UIUC-1. It points out that our approach can successfully retrieve the corresponding depth and skeleton features for actions to be recognized, and leverage the expanded features to achieve much better performance. In addition, our approach with the aid of cross-modal feature association, either considerably outperforms or is comparable to the state-of-the-art systems in each of the three benchmarks.

We investigate why the augmented depth and skeleton features help to improve recognition rates. Recall that the main difference between our approach and baseline RGB is that the augmented features are taken into account by the former, but ignored by the latter. To gain insight into their quantitative results, the confusion tables by the two approaches on the three benchmarks are shown in Figure 4. By comparing their accuracies class by class, we find out an interesting conclusion: The augmented depth and skeleton are particularly helpful in the actions that are characterized by small, local parts of human bodies, such as `check-watch`, `cross-arms`, `wave`, and `clapping`. In these actions, information captured by the RGB videos is limited, so borrowing features across modalities works.

7. Conclusions

The RGB-D cameras provide the opportunity for better solving increasingly complex computer vision tasks. However, the short effective distances are currently hindering their applicability. We have presented an approach to addressing the problem in this work. Our approach can borrow information from an offline collected, multi-modal database, and online augment actions with addition depth and skeleton features. Hence, it provides an alternative way of using RGB-D cameras even when they are not online accessible. Promising experimental results demonstrate that our approach can effectively adapt the variations between different databases, associate features across video modalities, and lead to remarkable boost in recognition performance. The proposed approach performs cross-modal feature association in a general way. We will generalize and apply this work to dealing with new vision applications where the emerging cameras are appreciated, such as borrowing features from binocular cameras for stereo vision applications, or from infrared cameras for night vision applications.

Acknowledgments. This work was supported by grants NSC 102-2221-E-001-025 and 100-2221-E-001-013-MY3.

References

[1] F. Bach, G. Lanckriet, and M. Jordan. Multiple kernel learning, conic duality, and the SMO algorithm. In *ICML*, 2004.

[2] O. Barnich and M. V. Droogenbroeck. Vibe: A universal background subtraction algorithm for video sequences. *TIP*, 2011.

[3] D. Cai, X. He, and J. Han. Semi-supervised discriminant analysis. In *ICCV*, 2007.

[4] J.-F. Cai, E. Candes, and Z. Shen. A singular value thresholding algorithm for matrix completion. *SIAM J. on Optimization*, 2010.

[5] L. Cao, Z. Liu, and T. Huang. Cross-dataset action detection. In *CVPR*, 2010.

[6] C. Chen and J. Aggarwal. Modeling human activities as speech. In *CVPR*, 2011.

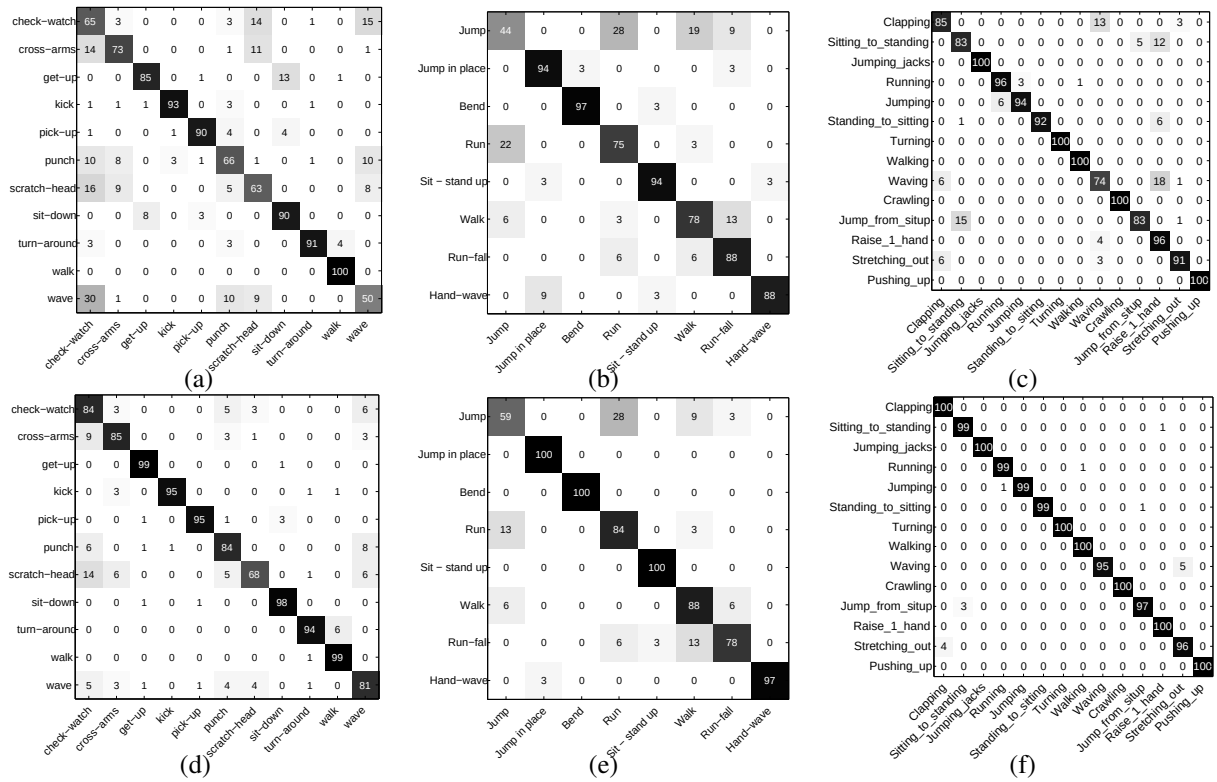


Figure 4. The confusion tables by two approaches, baseline RGB and ours, on the three benchmarks. (a) Baseline RGB on IXMAS. (b) Baseline RGB on i3Dpost. (c) Baseline RGB on UIUC-1. (d) Ours on IXMAS. (e) Ours on i3DPost. (f) Ours on UIUC-1.

[7] C.-Y. Chen and K. Grauman. Watching unlabeled video helps learn new human actions from very few labeled snapshots. In *CVPR*, 2013.

[8] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *CVPR*, 2005.

[9] R. Girshick, J. Shotton, P. Kohli, A. Criminisi, and A. Fitzgibbon. Efficient regression of general-activity human poses from depth images. In *ICCV*, 2011.

[10] N. Gkalelis, H. Kim, A. Hilton, N. Nikolaidis, and I. Pitas. The i3DPost multi-view and 3d human action/interaction database. In *Proc. Conf. Visual Media Production*, pages 159–168, 2009.

[11] J. Hernandez, R. Cabido, A. S. Montemayor, and J. Pantrigo. Human activity recognition based on kinematic features. *Expert Systems*, 2013.

[12] A. Iosifidis, A. Tefas, and I. Pitas. View-invariant action recognition based on artificial neural networks. *TNNLS*, 2012.

[13] I.-H. Jhuo, D. Liu, D. Lee, and S.-F. Chang. Robust visual domain adaptation with low-rank reconstruction. In *CVPR*, 2012.

[14] I. Laptev and T. Linderberg. Space-time interest points. In *ICCV*, 2003.

[15] Z. Lin, M. Chen, and Y. Ma. The augmented lagrange multiplier method for exact recovery of corrupted low-rank matrices. Technical report, UIUC, 2009.

[16] J. Liu, S. Ji, and J. Ye. Multi-task feature learning via efficient l2,1-norm minimization. In *UAI*, 2009.

[17] S. Maji, L. Bourdev, and J. Malik. Action recognition from a distributed representation of pose and appearance. In *CVPR*, 2011.

[18] P. Matikainen, M. Hebert, and R. Sukthankar. Representing pairwise spatial and temporal relations for action recognition. In *ECCV*, 2010.

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

[20] S. Pan and Q. Yang. A survey on transfer learning. *TKDE*, 2010.

[21] R. Poppe. A survey on vision-based human action recognition. *IVC*, 2010.

[22] K. Prabhakar, S. Oh, P. Wang, G. D. Abowd, and J. M. Rehg. Temporal causality for the analysis of visual events. In *CVPR*, 2010.

[23] A. Rakotomamonjy, F. Bach, S. Canu, and Y. Grandvalet. SimpleMKL. *JMLR*, 2008.

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

[25] N. C. Tang, C.-T. Hsu, C.-W. Su, T. K. Shih, and H.-Y. M. Liao. Video inpainting on digitized vintage films via maintaining spatiotemporal continuity. *TMM*, 2011.

[26] D. Tran and A. Sorokin. Human activity recognition with metric learning. In *ECCV*, pages 548–561, 2008.

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

[28] L. Wang and D. Suter. Recognizing human activities from silhouettes: motion subspace and factorial discriminative graphical model. In *CVPR*, 2007.

[29] D. Weinland, E. Boyer, and R. Ronfard. Action recognition from arbitrary views using 3D exemplars. In *ICCV*, 2007.

[30] D. Weinland, M. Özuysal, and P. Fua. Making action recognition robust to occlusions and viewpoint changes. In *ECCV*, 2010.

[31] Z. Wen and W. Yin. A feasible method for optimization with orthogonality constraints. Technical report, Rice University, 2010.

[32] X. Wu, D. Xu, L. Duan, and J. Luo. Action recognition using context and appearance distribution features. In *CVPR*, 2011.

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

[34] Q. Yin, X. Tang, and J. Sun. An associate-predict model for face recognition. In *CVPR*, 2011.

[35] G. Zhao and M. Pietikainen. Dynamic texture recognition using local binary patterns with an application to facial expressions. *TPAMI*, 2007.