Robust Action Recognition via Borrowing Information across Video Modalities

Nick C. Tang, Yen-Yu Lin, *Member, IEEE,* Ju-Hsuan Hua, Shih-En Wei, Ming-Fang Weng, and Hong-Yuan Mark Liao, *Fellow, IEEE*

***Abstract*—The recent advances in imaging devices have opened the opportunity of better solving the tasks of video content analysis and understanding. Next-generation cameras, such as the depth or binocular cameras, capture diverse information, and complement the conventional 2D RGB cameras. Thus, investigating the yielded multi-modal videos generally facilitates the accomplishment of related applications. However, the limitations of the emerging cameras, such as short effective distances, expensive costs, or long response time, degrade their applicability, and currently make these devices not online accessible in practical use. In this work, we provide an alternative scenario to address this problem, and illustrate it with the task of recognizing human actions. Specifically, we aim at improving the accuracy of action recognition in RGB videos with the aid of one additional RGB-D camera. Since RGB-D cameras, such as Kinect, are typically not applicable in a surveillance system due to its short effective distance, we instead offline collect a database, in which not only the RGB videos but also the depth maps and the skeleton data of actions are available jointly. The proposed approach can adapt the inter-database variations, and activate the borrowing of visual knowledge across different video modalities. Each action to be recognized in RGB representation is then augmented with the borrowed depth and skeleton features. Our approach is comprehensively evaluated on five benchmark datasets of action recognition. The promising results manifest that the borrowed information leads to remarkable boost in recognition accuracy.**

***Index Terms*—Action recognition, next-generation cameras, transfer learning, feature borrowing**

I. Introduction

In the past decade, human action recognition had become one of the most important research topics in video content analysis and understanding. A vast amount of research effort had been made to establish representative benchmarks and propose effective recognition schemes. Despite the great effort, action recognition in general is still very challenging, and most action recognition systems suffer from the difficulties caused by large *intra-class variations* [1]. Although designing more powerful features from RGB videos has gained significant progress, the information captured by conventional RGB cameras is insufficient to account for various types of the unfavorable variations, such as those caused by mutual or self-occlusion, camera perspective settings, and inter-personal differences.

Most video processing techniques are highly adapted to the imaging devices. We are aware of the recent advances in imaging devices, such as the RGB-D camera Microsoft Kinect [2], the binocular camera FUJIFILM FinePix Real 3D [3], the infrared camera FLIR T620 [4], and the lightfield camera Lytro [5]. The multi-modal videos they record give rich and diverse information. Thus, there has been a strong demand for content analysis techniques that leverage these cameras to better solve increasingly complex video processing tasks including action recognition, and even to initiate new applications. However, these cameras have their respective restrictions. For instance, Kinect is with a short range of effective distance from 1.2 to 3.5 meters, and the emerging cameras are often relatively expensive to conventional RGB cameras. The restrictions hinder their applicability, and may make these devices not online accessible in practical use.

In this work, we propose an alternative scenario to address the foregoing problem, and focus on boosting the performance of the underlying applications by jointly using a conventional 2D RGB camera and one additional emerging camera, even if the latter is not online available. We illustrate this scenario with the application to recognizing human actions, a fundamental topic in video processing.

As pointed out in [6], [7], the depth maps taken by an RGB-D camera as well as the inferred skeleton data of human actions are very helpful toward more accurate action recognition. However, most RGB-D cameras, such as Kinect, are not applicable in video surveillance systems due to the short effective distance. We instead use Kinect to offline collect an auxiliary, multi-modal database that contains entries in form of triplets: the RGB videos as well as the depth maps and the skeleton structures of action instances. Our goal is to improve the performance of recognizing actions taken by a 2D RGB camera by leveraging the knowledge borrowed from the auxiliary database. More specifically, the proposed approach considers the action to be recognized as a query to the auxiliary database, and attempts to retrieve the corresponding depth map and skeleton structure. If it works, it compensates for the online unavailability of Kinect.

Fig. 1 outlines the framework. The proposed approach to cross-modal information borrowing is composed of three stages. At the first stage, our approach attempts to establish the

This work was supported in part by Ministry of Science and Technology (MOST) under Grant 103-2221-E-001-026-MY2 and Grant 103-2221-E-001-009-MY3.

N. C. Tang is with the Institute of Information Science, Academia Sinica, Taipei 115, Taiwan and the Department of Information and Computer Engineering, Chung-Yuan University, Chung Li 320, Taiwan. E-mail: nickctang@gmail.com.

Y.-Y. Lin is with the Research Center for Information Technology Innovation, Academia Sinica, Taipei 115, Taiwan. E-mail: yylin@citi.sinica.edu.tw.

J.-H. Hua and S.-E. Wei are with the Robotics Institute, Carnegie Mellon University, Pittsburgh PA 15213, USA. E-mail: rushaneh@cs.cmu.edu and shihenw@cmu.edu.

M.-F. Weng is with the Smart Network System Institute, Institute for Information Industry, Taipei City 105, Taiwan. E-mail: mfueng@iii.org.tw.

H.-Y. M. Liao is with the Institute of Information Science, Academia Sinica, Taipei 115, Taiwan. E-mail: liao@iis.sinica.edu.tw.

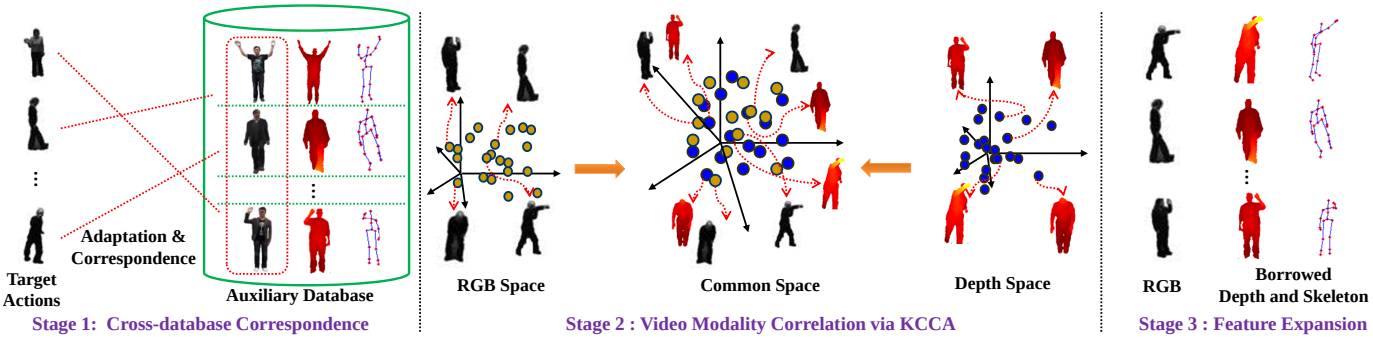


Fig. 1. Our approach aims at improving action recognition by borrowing information from an auxiliary, offline collected database where multi-modal videos are available. To this end, it establishes cross-database correspondences, correlates domains of different video modalities, and associates each action instance in RGB representation with the reconstructed depth and skeleton features. In this manner, our approach provides an alternative way of utilizing new types of cameras even if they are not online accessible.

correspondences between the RGB data in the target database and the auxiliary database. It adapts to the variations of the two databases by formulating the task of data correspondences as a labeling problem over *Markov random fields* (MRFs) [8]. In this way, both the data appearances and the label information can be simultaneously taken into account, and result in better cross-database correspondences. The second stage is built upon the correspondences. We correlate videos of different modalities by adopting *kernel canonical correlation analysis* (KCCA) [9], which projects multi-modal videos in the two databases into a common space. In that space, the actions in the target database can then be convexly reconstructed by the nearby entries in the auxiliary database. At the third stage, we consider the actions to be recognized as the queries with their reconstructions in different modalities as returns, the actions are then augmented with borrowed depth and skeleton features. It follows that techniques, like early and late fusion, can be adopted to explore the complementary information carried by the original and the borrowed features, and lead to performance improvement in action recognition.

The main contribution of this work is to provide an effective way of utilizing new types of cameras, and better solve complex applications even when these cameras are not online available. The proposed approach is comprehensively evaluated on five benchmarks of action recognition, each of which is established for addressing specific issues and contains actions of different classes. By using the same auxiliary database, our approach results in remarkable accuracy improvement in each dataset. It validates the robustness and flexibility of the proposed approach. Furthermore, our approach is developed in a general manner, and hence can be applied to other applications in which multi-modal videos are helpful, such as gesture recognition or anomaly detection.

II. Related Work

In the section, we review some research topics that are relevant to the proposed framework, including action recognition, transfer learning, and heterogeneous feature fusion.

A. Action Recognition

Human action recognition has received strong attention in the fields of computer vision and video processing. Being one of the most important components in video understanding, action recognition is essential to widespread applications, such as surveillance and human-computer interaction. As indicated in [1], one fundamental difficulty of action recognition is the *large intra-class variations*. These variations can result from both intrinsic and extrinsic factors, such as posture differences among subjects, clutter background, different camera perspectives, mutual or self occlusions.

To account for intra-class variations, many feature descriptors have been proposed to better characterize actions. *Global descriptors*, e.g., [10], [11], which characterize and encode the region of an action as a whole, are popular for their simplicity. For instance, Bobick and Davis [10] extracted the silhouettes of an action, and recognize the action by analyzing the motions of the silhouettes. Gorelick et al. [11] represented an action as a space-time shape and adopted Poisson equation to extract the space-time features for classification. However, global descriptors are often sensitive to occlusions and deformations. On the other hand, *local descriptors*, especially the *bag-of-words* models [12], [13], are widely used recently. Approaches of this class typically compile histograms of locally quantized features. However, the geometric structure among local features in the spatio-temporal space is ignored in these approaches, possibly resulting in performance degradation.

To address this issue, one of the current research trends in action recognition is to model the relationships among local features. For example, Matikainen et al. [14] specified the geometrical displacements between local features by generating a frequency lookup table. Prabhaka et al. [15] computed the causalities between visual words, and included them as parts of the features. Besides, graphical models, such as *factorial conditional random fields* in [16] or *hidden Markov model* in [17], have been applied to formulate the spatio-temporal correlation of local evidences. All the aforementioned methods recognized actions based on 2D RGB images/videos. Restricted by the available information, it is still very challenging to deal with intra-class variations caused by different camera perspectives or partial occlusions.

Owing to the recent advances in sensor technology, it has been feasible to capture color as well as depth information of an action video in real time by RGB-D cameras, e.g., Kinect.

Research efforts, such as [6], [18], [19], have demonstrated that depth maps of actions afford informative and invariant clues to build robust action recognition or pose estimation systems. Besides, *OpenNI library* [20] 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., [7], [21], on 3D skeleton representation and correction open the opportunity of resolving multi-view action recognition. In [22], Ellis et al. utilized skeletons of human actions to perform both pre-segmented and online action recognition. Ashraft et al. [23] suggested to represent an action as a set of projective depths with respect to planes extracted from mirror symmetry, and carried out view-invariant action recognition. The introduction of depth and skeleton information indeed benefits action recognition. However, the short ranges of the effective distances make RGB-D cameras inapplicable in many real world applications, such as surveillance where the installed cameras are usually distant from the monitored environments. In our prior work [24], an approach to augmenting additional depth and skeleton features to a target action was proposed. In this work, we generalize our approach to dealing with actions that are not fully covered by an auxiliary dataset, and comprehensively evaluate it on five benchmarks of action recognition.

B. Transfer Learning

Transfer learning refers to a knowledge delivering process from the *source* domains to the *target* domain. It aims to help improve the task in the target domain by leveraging abundant information in the source domains. The soul of transfer learning is to identify the domain-specific and the commonly-shared knowledge in the sources, and transfer the latter to benefit target task. According to the survey paper [25], the methods of knowledge transfer can be generally divided into four categories: transfer by *model parameters* [26], by *data instances* [27], [28], by *feature representation* [29], and by *relational information* [30].

Most of the above-mentioned methods work when the source and target domains are the same or related. In our case, the source and the target domains correspond to different video modalities, and are hence irrelevant. To handle this issue, we adopt kernel CCA to correlate video data of different modalities, and uncover a common subspace, upon which knowledge transfer across modalities is allowed.

C. Heterogeneous Feature Fusion

Two popular strategies to fuse heterogeneous features are *early fusion* and *late fusion* [31]. While the former fuses information in the level of features, the latter combines the predictions of the models, each of which is derived with one or a subset of features.

One representative way of early fusion is *multiple kernel learning* (MKL) [32]–[34], which refers to learning a kernel machine with multiple kernels. Recent research efforts [35], [36] have shown that fusing feature in the kernel space not only increases the accuracy but also enhances the interpretability of the yielded classifiers. In our case, we could represent data described by each type of the original and the borrowed features as a kernel matrix. MKL will learn a kernel machine and derive the kernel weights. Namely, the heterogeneous features are combined in the domain of kernel matrices.

In late fusion, one additional classifier or regressor is typically employed to merge the confidence scores of the models separately constructed from different features. Late fusion, e.g., [30], [31], is relatively easy to implement, but still shows effectiveness in practice. We apply both one early fusion method and one late fusion method to fuse the multi-modal features in our case. The two methods achieve similar performance in the aspect of accuracy. However, higher computational cost is required in the early fusion method for searching the optimal values of the hyperparameters in the kernel functions.

III. Problem Statement

We focus on recognizing actions of C classes. Suppose that we are given a training set, $D = \{(\mathbf{x}_i, y_i)\}_{i=1}^{N}$, where $\mathbf{x}_i \in \mathcal{X}$ and $y_i \in \mathcal{Y} = \{1, 2, ..., C\}$ are the RGB feature representation and the class label of the ith action, respectively. To enhance the performance of action recognition, an auxiliary dataset, $A = \{(\tilde{\mathbf{x}}_i, \tilde{\mathbf{d}}_i, \tilde{\mathbf{s}}_i)\}_{i=1}^{M}$, taken by Kinect is also provided, 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 feature representations of the ith instance, respectively. Note that auxiliary dataset A is unlabeled, and we use *tildes* to mark data in A for the sake of clearness. We will utilize information in D and A, and derive a better classifier for predicting test data that are similarly distributed to D. More specifically, we consider actions in D as queries to A, and focus on retrieving their corresponding depth maps and skeleton data. That is, we aim at leveraging the RGB-D camera Kinect even though it is not online accessible in the application.

The auxiliary dataset A we collected was compiled to cover the action classes of interest in advance, i.e., $\mathcal{Y}$, in this case. Establishing A beforehand is reasonable, since we often focus on detecting some predefined types of actions in most action recognition applications. However, it is not necessary that the action classes in D and that in A are the same. For instance, D in turn is one of the adopted benchmarks of action recognition in our experiments, and the action classes in A are the union of those in all the benchmarks. In addition, D and A are allowed to be established in different ways, so large inter-database variations may be caused.

IV. The Proposed Approach

Our approach improves action recognition by augmenting each action, initially in RGB feature presentation, with the estimated depth and skeleton features. It is composed of three components, each of which is described in turn as follows.

A. Cross-database Correspondences

The goal of this stage is to correlate D and A, the two independently collected datasets, by exploring their common video modality, RGB. Specifically, we want to associate each $\mathbf{x}_i$ in D with a plausible sample $\tilde{\mathbf{x}}_{\pi_i}$ in A. A naïve way is the nearest neighbor search. However, it ignores the inter-database variations, and may result in sub-optimal performance. To address this issue, we exploit the data labels in D, and incorporate discriminant analysis to guide the construction of

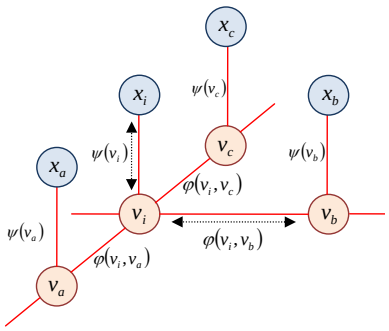


Fig. 2. The illustration figure of our MRF model. We construct a state node and an observation node for each action $\mathbf{x}_i \in D$, and connect the two nodes. The state node of $\mathbf{x}_i$ is further connected to the state nodes of the ℓ nearest neighbors of $\mathbf{x}_i$, i.e., $\mathbf{x}_a$, $\mathbf{x}_b$, and $\mathbf{x}_c$ here. Two types of energy functions, ψ and φ, defined over the edges are considered.

cross-database correspondences. We cast this task as a labeling problem over Markov random fields (MRFs), in which the mutual verification among correspondences is activated. Hence, the borrowed multi-modal features are more discriminative.

In the construction of the MRFs model with graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, each $\tilde{\mathbf{x}}_i$ in A corresponds to a *state*, while each $\mathbf{x}_i$ in D is associated with a variable node v_i. There are total M states and N variable nodes. Let $\mathcal{L} = \{1, 2, ..., M\}$ denote the set of the states. Each node $v_i \in \mathcal{V}$ takes a value from $\mathcal{L}$. In this way, v_i determines the correspondence of $\mathbf{x}_i$ in A, i.e., $\{\mathbf{x}_i \in D, \tilde{\mathbf{x}}_{v_i} \in A\}$. An undirected edge $e = (v_i, v_j)$ is added into $\mathcal{E}$ if $\mathbf{x}_j$ is one of the ℓ nearest neighbors of $\mathbf{x}_i$. Namely, $|\mathcal{L}| = M$, $|\mathcal{V}| = N$, and $\ell N/2 \leq |\mathcal{E}| \leq \ell N$. Here we use Euclidean distance for the nearest neighbors search. Suppose the average number of training data per class is n. The value of ℓ is empirically set as $\lceil \sqrt{n} \rceil$ here. MRFs model the probability distribution over each possible labeling $V = [v_1 \cdots v_N] \in \mathcal{L}^N$ in form of

$$P(V) = \frac{1}{Z} \exp\left(-E(V)\right), \tag{1}$$

where *partition function* Z for normalization is defined as

$$Z = \sum_{V' \in \mathcal{L}^N} \exp\left(-E(V')\right). \tag{2}$$

MRFs allow the flexibility of designing proper energy function E according to our prior knowledge about the problem.

In this work, we consider the following *energy function*:

$$E(V) = \sum_{v_i \in \mathcal{V}} \psi(v_i) + \sum_{(v_i, v_j) \in \mathcal{E}} \varphi(v_i, v_j), \tag{3}$$

where the *unary function* ψ and the *pairwise function* φ are respectively defined as

$$\psi(v_i) = \begin{cases} \|\mathbf{x}_i - \tilde{\mathbf{x}}_{v_i}\|, & \text{if } \tilde{\mathbf{x}}_{v_i} \in k\text{NNs of } \mathbf{x}_i \text{ in } A, \\ \infty, & \text{otherwise,} \end{cases} \tag{4}$$

$$\varphi(v_i, v_j) = \begin{cases} \lambda_1 \|\tilde{\mathbf{x}}_{v_i} - \tilde{\mathbf{x}}_{v_j}\|, & \text{if } y_i = y_j, \\ -\lambda_2 \|\tilde{\mathbf{x}}_{v_i} - \tilde{\mathbf{x}}_{v_j}\|, & \text{otherwise,} \end{cases} \tag{5}$$

where kNNs denote the k nearest neighbors. k, λ_1, and λ_2 are three positive constants. Their values are determined by cross validation in the experiments. An illustration figure of our MRF model is shown in Fig. 2.

The designed unary function in (4) prefers a high degree of similarity between $\mathbf{x}_i$ and $\tilde{\mathbf{x}}_{v_i}$, and hence ensures the appearance compatibility of each correspondence. The pairwise function in (5) enforces class-consistent labeling. That is, $\tilde{\mathbf{x}}_{v_i}$ and $\tilde{\mathbf{x}}_{v_j}$ are required to be similar to each other if and only if $\mathbf{x}_i$ and $\mathbf{x}_j$ are of the same class.

After applying *graph cut* [37] to minimizing the energy in (3), the most plausible configuration V is obtained. It follows that the N cross-database correspondences, $\{(\mathbf{x}_i, \tilde{\mathbf{x}}_{\pi_i})\}_{i=1}^N$, are established with $\pi_i \leftarrow v_i$.

It is worth mentioning that the labels of training actions are taken into account in the MRF model. The corresponding actions in A of the actions in D of the same class tend to be similar, and to be dissimilar otherwise. It implies that more discriminant depth and skeleton features can be borrowed in the successive stages.

B. Cross-modal Feature Association

At the stage, we aim to augment each training action in D and each testing action with additional depth and skeleton features. Based upon the one-to-one modal mapping in A, the correspondences $\{(\mathbf{x}_i, \tilde{\mathbf{x}}_{\pi_i})\}_{i=1}^N$ established above can be propagated across video modalities, i.e., $\{(\mathbf{x}_i, \tilde{\mathbf{d}}_{\pi_i})\}_{i=1}^N$ and $\{(\mathbf{x}_i, \tilde{\mathbf{s}}_{\pi_i})\}_{i=1}^N$. Yet, these correspondences are valid only for training data in D, and are not available for testing data. To overcome this problem, we adopt *kernel canonical correlation analysis* (KCCA) to correlate data of two different domains, RGB $\mathcal{X}$ and skeleton $\mathcal{S}$, via $\{(\mathbf{x}_i \in \mathcal{X}, \tilde{\mathbf{s}}_{\pi_i} \in \mathcal{S})\}_{i=1}^N$.

Let $\phi : \mathcal{X} \to \mathcal{F}_x$ denote the feature map, which transforms data from domain $\mathcal{X}$ to space $\mathcal{F}_x$. Similarly, we have $\tilde{\phi} : \mathcal{S} \to \mathcal{F}_s$. Via ϕ and $\tilde{\phi}$, data of the two domains are mapped to high-dimensional Hilbert spaces, i.e.,

$$\mathbf{x}_i \mapsto \phi(\mathbf{x}_i) \quad \text{and} \quad \tilde{\mathbf{s}}_i \mapsto \tilde{\phi}(\tilde{\mathbf{s}}_i), \text{ for } i = 1, 2, ..., N. \tag{6}$$

KCCA seeks a pair of projections $(\mathbf{u}, \mathbf{v})$ to uncover a common space, in which the correlation between projected data $\{\mathbf{u}^\top \phi(\mathbf{x}_i)\}$ and $\{\mathbf{v}^\top \tilde{\phi}(\tilde{\mathbf{s}}_{\pi_i})\}$ is maximized. It has been proven in [38] that the projections lie in the span of data, i.e.,

$$\mathbf{u} = \sum_{i=1}^N \alpha_i \phi(\mathbf{x}_i) \quad \text{and} \quad \mathbf{v} = \sum_{i=1}^N \beta_i \tilde{\phi}(\tilde{\mathbf{s}}_{\pi_i}). \tag{7}$$

In KCCA, the optimal projections $(\mathbf{u}^*, \mathbf{v}^*)$, parameterized by $(\boldsymbol{\alpha}^* = [\alpha_1^* \cdots \alpha_N^*]^\top, \boldsymbol{\beta}^* = [\beta_1^* \cdots \beta_N^*]^\top)$, are given by

$$(\boldsymbol{\alpha}^*, \boldsymbol{\beta}^*) = \arg\max_{\boldsymbol{\alpha}, \boldsymbol{\beta}} \frac{\boldsymbol{\alpha}^\top K_x K_s \boldsymbol{\beta}}{\sqrt{\boldsymbol{\alpha}^\top K_x^2 \boldsymbol{\alpha} \cdot \boldsymbol{\beta}^\top K_s^2 \boldsymbol{\beta}}}, \tag{8}$$

$$\text{where } K_x = [\phi(\mathbf{x}_i)^\top \phi(\mathbf{x}_j)] \in \mathbb{R}^{N \times N}, \tag{9}$$

$$K_s = [\tilde{\phi}(\tilde{\mathbf{s}}_{\pi_i})^\top \tilde{\phi}(\tilde{\mathbf{s}}_{\pi_j})] \in \mathbb{R}^{N \times N}. \tag{10}$$

It can be verified that the optimal $(\boldsymbol{\alpha}^*, \boldsymbol{\beta}^*)$ in (8) can be obtained by solving a generalized eigenvalue problem. Furthermore, the formulation of KCCA can be generalized to uncover multidimensional projections, i.e., $U = [\mathbf{u}_1 \cdots \mathbf{u}_p]$ and $V = [\mathbf{v}_1 \cdots \mathbf{v}_p]$. In implementation, we use the RBF kernel functions for implicitly computing the inner products in (9)

and (10). Namely,

$$K_x(i,j) = \phi(\mathbf{x}_i)^\top \phi(\mathbf{x}_j) = \exp\left(\frac{-\|\mathbf{x}_i - \mathbf{x}_j\|^2}{\sigma_x^2}\right) \text{ and} \quad (11)$$

$$K_s(i,j) = \tilde{\phi}(\tilde{\mathbf{s}}_{\pi_i})^\top \tilde{\phi}(\tilde{\mathbf{s}}_{\pi_j}) = \exp\left(\frac{-\|\tilde{\mathbf{s}}_{\pi_i} - \tilde{\mathbf{s}}_{\pi_j}\|^2}{\sigma_s^2}\right), \quad (12)$$

where σ_x and σ_s are positive constants. As suggested in [39], we set σ_x as the mean of the pairwise distances among data $\{\mathbf{x}_i\}_{i=1}^N$. The same way is applied to determining σ_s.

With U and V, we first project all the data under skeleton representation in A, i.e., $\{V^\top \tilde{\mathbf{s}}_i\}_{i=1}^M$. For an input action $\mathbf{x}$, which is either a training or a testing sample, we project $\mathbf{x}$ by $U^\top \mathbf{x}$, and retrieve its m nearest skeleton samples. Without loss of generality, we assume that the retrieved samples are $\{V^\top \tilde{\mathbf{s}}_i\}_{i=1}^m$. The borrowed skeleton feature $\mathbf{s}$ can be generated by minimizing the square reconstruction error, and is a convex combination of the m retrieved samples. That is,

$$\mathbf{s} = \sum_{i=1}^m \gamma_i \tilde{\mathbf{s}}_i, \quad (13)$$

where the coefficients $\{\gamma_i\}_{i=1}^m$ are determined by solving

$$\min_{\{\gamma_i\}_{i=1}^m} \quad \|U^\top \mathbf{x} - \sum_{i=1}^m \gamma_i V^\top \tilde{\mathbf{s}}_i\|^2, \quad (14)$$

$$\text{s.t.} \quad \gamma_i \geq 0 \text{ for } i = 1, 2, ..., m, \quad (15)$$

$$\sum_{i=1}^m \gamma_i = 1. \quad (16)$$

The constrained least square problem in (14) can be efficiently solved by using the algorithm suggested in [40]. We tune the value of m via cross validation. The optimal range of m is $1 \sim 5$ in most of our experiments.

The same procedure is repeated for correlating modalities RGB $\mathcal{X}$ and depth map $\mathcal{D}$, and the depth map, $\mathbf{d}$, of $\mathbf{x}$ is similarly retrieved. Action $\mathbf{x}$ is then augmented with two additional features borrowed from auxiliary database A:

$$\mathbf{x} \mapsto (\mathbf{x}, \mathbf{d}, \mathbf{s}). \quad (17)$$

The strategy is applied to each sample in the target database D. It follows that the *augmented dataset* is constructed, i.e., $D' = \{(\mathbf{x}_i, \mathbf{d}_i, \mathbf{s}_i)\}_{i=1}^N$.

Fig. 3 gives an example of the feature augmentation. Fig. 3(a) shows the action to be recognized, the query. The top three retrieved depth maps and the skeleton structures, i.e., those with largest reconstruction coefficients in (13), are shown in Fig. 3(c) and Fig. 3(d), respectively. We also give the corresponding RGB videos of the three depth maps in Fig. 3(b) for visually assessing the similarity between the query and the returns.

C. Recognition with The Aid of The Augmented Features

By treating actions in the target database as queries to the auxiliary database, their corresponding depth maps and skeleton structures are retrieved by the aforementioned procedure. The proposed approach implements the principle of *query expansion* in the sense that the seed query is reformulated and expanded to improve the performance of the subsequent applications. Unlike most previous approaches to query expansion, ours carries out the feature expansion across video modalities captured by different types of cameras. It results in more complementary information that can be leveraged to facilitate the accomplishment of action recognition.

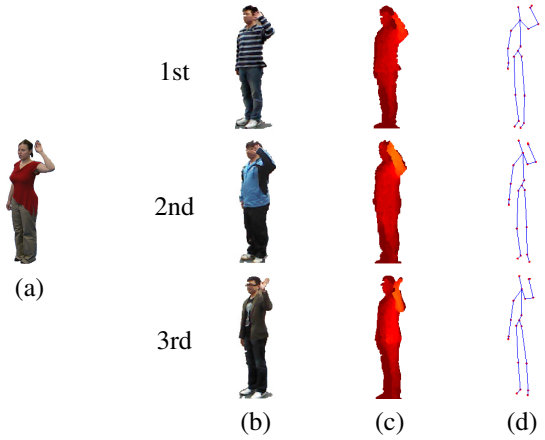


Fig. 3. Feature augmentation. (a) An action in the target database. (c) The top three retrieved depth maps. (b) The corresponding RGB videos of the depth maps. (d) The top three retrieved skeleton structures.

The training data have been expanded from D to D'. Three video modalities of each action are available at the same time. Early fusion or late fusion can be adopted for combining the three heterogeneous features to achieve better performance. We have implemented both the two fusion strategies, and describe them in the following.

1) Multiple Kernel Learning for Early Fusion

We compile an kernel matrix for actions in each of the three image modalities, and adopt *SimpleMKL* [33], one of the state-of-the-art MKL packages, to learn an SVM classifier with multiple kernels. In this way, the three heterogeneous features are fused in the domain of kernel matrices.

2) Top-level Logistic Regression for Late Fusion

We learn an SVM classifier with probability estimation for data in each modality, and concatenate the outputs of all the SVM classifiers. A top level L_2-regularized logistic regressor is derived to work on the concatenated vectors for feature fusion. In this manner, features are combined in the classifier level.

D. On Predicting A Test Action

Given a test action $\mathbf{x}$, we first augment it with the borrowed depth and skeleton features via (17). Then, either early fusion or later fusion can be applied to completing the prediction.

Testing with the benchmarks of action recognition, the performances of early fusion and late fusion are quite similar. Multiple kernel learning is less efficient owing to jointly tuning the hyperparameters in the kernel functions. Thus, we choose late fusion, and will report quantitative results by late fusion in all the experiments.

V. Video Feature Representation

In the section, we describe the adopted features for characterizing actions in RGB videos, depth maps, and skeleton structures, respectively.

A. Features for RGB Videos

We implement two state-of-the-art methods to extract robust RGB features from action videos that contain *static* background and *dynamic* background respectively.

For action videos with static background, we preprocess each video as follows. First, we apply the video inpainting technique [41] to compute the background images from a collection of sample videos. Then, we take the acquired background images as the mask, and adopt a background subtraction algorithm [42] to segment the foreground region in each video frame. Accordingly, we can precisely compute the space-time volume (STV) features from the region of interest without worrying about the cluttered background. In our implementation, we scale down a given action video to the resolution $48 \times 64 \times t$, where t is the number of frames in the video. The 3*D-HOG* (histogram of oriented gradients) descriptor [43] is applied to extract features both in a space-time volume and its horizontal mirror for against reflection. In more detail, we use $16 \times 16 \times 16$ pixel blocks, each of which is further divided into $2 \times 2 \times 2$ cells. Five hundred prototypes are derived to build up the embedding space. It leads to a compact representation for actions in RGB videos. In our experiments, the RGB features are used to describe actions in IXMAS, i3DPost, and UIUC1 datasets.

We adopt the robust method proposed in [44] to extract RGB features for action videos with dynamic background. In [44], dense points are sampled from each frame, and are tracked based on displacement information from a dense optical flow field. Refer to [44] for the details of implementation. In our experiments, the RGB features are used to describe actions in UCF-CIL and UCF human action datasets.

B. Features for Depth Maps

We are motivated by the good performance reported in [45], and use the *Spatio-temporal Local Binary Pattern* (STLBP) as the feature representation of depth maps. The STLBP is developed to model the variation of motion and appearance based on concatenated LBP histograms. In our implementation, we first compute the LBP histograms of each depth map, and then extract the co-occurrence features from neighboring points on the three orthogonal planes. Refer to [45] for the details of implementation.

C. Features for Skeleton Structures

As for skeleton features, we employ the *Fourier Temporal Pyramid* [21] to represent the temporal dynamics of each 3D joint of a human body. Computing the temporal dynamics involves the following steps: First, pairwise relative position features are extracted for each joint. On the one hand, enumerating the difference between all the joint pairs yields a redundant representation, which increases discriminative power because actions are generally interpreted by considering contexts. On the other hand, some joints in the frame could be outliers. Thus, we perform PCA on the skeletal structures of all training frames to learn a compact representation which is more robust to noise. Then, by tracking the changes of feature values, each element of the representation can be considered as a set of time series data. For learning a representation which is insensitive to temporal misalignment, we use Wang et al.'s approach [21] to derive the Fourier Temporal Pyramid features to capture the temporal structure of an individual observation. In our current implementation, we use a three-level Fourier temporal pyramid with $1/4$ length of each segment as low-frequency coefficients.

VI. Experimental Results

To test the effectiveness of our approach, we present the performance of our approach to action recognition and compare it with other state-of-the-art methods. In Section VI-A, we shall describe the five used benchmarks of action recognition as well as the auxiliary, multi-modal database that we collected. The adopted baselines for performance comparison are introduced in Section VI-B. For performance evaluation, we first consider the cases where the target actions are covered by the auxiliary dataset, and show both the quantitative and qualitative results of our approach. Finally, we also consider the cases where the target actions are only partially covered by the auxiliary dataset, and discuss the obtained experimental results.

A. Action Recognition Benchmarks and Auxiliary Database

Five benchmarks of action recognition, including IXMAS [46], i3DPost [47], UIUC1 [48], UCF sports [49], and UCF CIL [50], are adopted in performance evaluation. They all contain RGB action videos captured by stationary camcorders. Since our method performs visual knowledge borrowing across distinct data modalities, we use Microsoft Kinect to build up an auxiliary, multi-modal action database. This three-modal dataset will serve as the common auxiliary database in all the experiments on the five benchmarks. Since the five benchmarks were designed for addressing specific issues and were compiled in different environment, not only the performance gain of feature borrowing but also the generalization of our approach are jointly assessed in this experiment setup.

The auxiliary dataset we collected as well as the five adopted benchmarks are described as follows.

- **Auxiliary Dataset:** We used Kinect to collect this dataset, so the RGB frames, the depth maps and the corresponding skeletons of each action sequence are available simultaneously. The auxiliary dataset is composed of 1600 action sequences of 40 distinct action classes. Total 10 actors were employed in the construction of the dataset. Each of them performed all the 40 classes of actions to cover all the action categories in the first three benchmarks (IXMAS, i3DPost, and UIUC1), and partially cover action categories in the last two benchmarks (UCF sports and UCF CIL). Each action was recorded by two cameras, respectively located with view angles of $0\,^\circ$ and $45\,^\circ$. Besides, we mirrored each recorded action for against reflection. Fig. 4 shows an action example from each category of the dataset. The example actions in the top two rows were taken by the camera with view angle of $0\,^\circ$, while the rest were by the camera with view angle of $45\,^\circ$.
- **IXMAS:** The IXAMS (INRIA XMAS Motion Acquisition Sequences) dataset [46] is composed of 12 action classes. They are: *check watch*, *cross arms*, *scratch head*, *sit down*, *get up*, *turn around*, *walk*, *wave*, *punch*, *kick*, *point* and

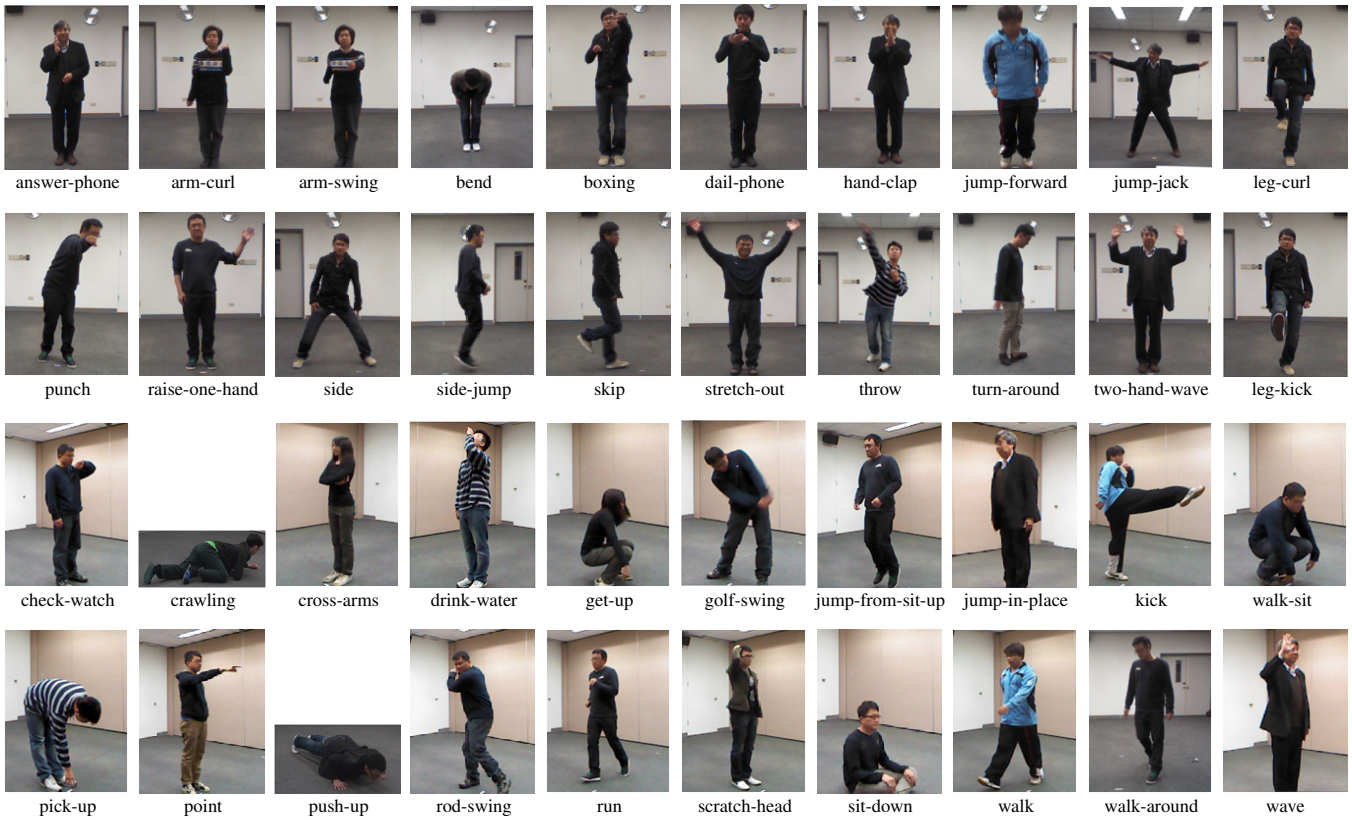


Fig. 4. The auxiliary dataset. One example comes from each of the 40 action classes. The examples in the top two rows were taken by the camera with view angle of $0\,^\circ$, while the rest were by the camera with view angle of $45\,^\circ$.

pick up. Each action was performed by 10 different actors for 3 times. All actions were recorded by 5 cameras with different viewing angles. The foreground mask used to locate a human body was provided in the dataset. Note that we conducted the comparisons on action videos captured by camera 1, 2 and 3 in the IXMAS dataset, since these videos were taken with the view angles close to either $0\,^\circ$, $45\,^\circ$, or $-45\,^\circ$. In addition, these action sequences have been used in the experiments conducted by other state-of-the-art methods, e.g., [51]–[53].

- **i3DPost:** The i3DPost Multi-view Human Action Dataset [47] contains 96 high-resolution video sequences of 12 action types performed by 8 actors. These actions were recorded by multiple cameras with 8 different viewpoints. Each of these cameras was arranged to have $45\,^\circ$ difference with its direct neighbors so that a full $360\,^\circ$ coverage can be achieved. The actions collected in this dataset include ten daily activities: *walk*, *run*, *jump-forward*, *pick-up*, *wave-right-hand*, *jump-in-place*, *sit*, *fall*, *walk-sit*, *run-jump-walk* and two human interactions: *handshake* and *pull*. The studio was covered by blue background. Thus, the foreground mask used to characterize a human's body can be extracted by jointly using video inpainting [41] and the background subtraction technique [42]. For comparing with other state-of-the-art methods, we followed the evaluation protocols suggested in [54], [55], in which total eight daily activities and three camera settings, include two single-view settings ($0\,^\circ$ and $45\,^\circ$, respectively), and their combination, were adopted in the experiments.
- **UIUC1:** The UIUC1 human activity dataset [48] is composed of 532 high resolution sequences of 14 kinds of activities performed by 8 actors. Each actor performed each activity 5 times. The activities performed in this dataset are *Walk*, *Run*, *Jump-upward*, *Wave*, *Jumping-jacks*, *Hand-clap*, *Jump-from-sit-up*, *Raise-one-hand*, *Stretch-out*, *Turn-around*, *Sit-to-stand*, *Crawl*, *Push-up* and *Stand-to-sit*. The foreground mask used to capture a human body is provided in the dataset.
- **UCF sports:** The UCF sports action dataset is composed of 150 video sequences which are collected from the Internet. It contains various sports actions. The collected actions include *swinging on the pommel*, *diving*, *kicking*, *weight-lifting*, *horse-riding*, *running*, *skateboarding*, *swinging-at the high bar*, *golf swinging*, and *walking*. The action videos were captured by different camera angles and of various resolutions. In addition, these actions were performed by different players. It means that large intra-class variations present in this dataset.
- **UCF CIL:** The UCF CIL dataset consists of 56 sequences of 8 actions, including *ballet fouettes*, *golf swing*, *push-up exercise*, *ballet spin*, *one-handed tennis backhand stroke*, *two-handed tennis backhand stroke*, *tennis-serve*, *tennis forehand stroke*. These action sequences were performed by different human subjects, and were taken by different cameras with diverse viewpoints. The ground truth of the skeletons is given in this dataset.

TABLE I
RECOGNITION RATES (%) OF DIFFERENT METHODS ON THE IXMAS DATASET.

Method	Ours	RGB	Bor-DEP	Bor-SKE	RGB+SDA	1NN-Bor	[52]	[53]
Accuracy (%)	91.6	89.4	85.3	84.8	90.0	78.8	87.7	81.3

TABLE II
RECOGNITION RATES (%) OF DIFFERENT METHODS ON THE I3DPOST DATASET.

Method		Ours	RGB	Bor-DEP	Bor-SKE	RGB+SDA	1NN-Bor	[54]
Accuracy (%)	0°	95.2	86.9	88.1	80.9	76.8	72.6	77.5
	45°	96.4	91.6	83.3	84.5	85.3	77.3	84.9
	$0^\circ \cup 45^\circ$	94.7	85.1	88.7	86.3	87.5	80.9	84.9

TABLE III
RECOGNITION RATES (%) OF DIFFERENT METHODS ON THE UIUC1 DATASET.

Method	Ours	RGB	Bor-DEP	Bor-SKE	RGB+SDA	1NN-Bor	[48]	[44]	[56]
Accuracy (%)	99.4	94.9	87.4	88.2	95.4	88.5	98.3	98.4	99.6

B. Baselines

For performance analysis and comparison, we implemented the following five baselines, each of which is denoted below in bold and in abbreviation:

- **RGB**: This baseline simply ignores the information from the auxiliary database. It extracts the RGB features, described in Section V, for the actions in the target database (one of the five adopted benchmarks), and employs an SVM classifier to make the prediction. Note that for experiments on IXMAS, i3DPost, and UIUC1 datasets, the 3D-HOG descriptor [43] is used to extract RGB features. For the experiments on UCF sports and UCF CIL datasets, the RGB features are computed by the method proposed in [44]. We can examine whether the auxiliary database helps improve the performance of action recognition by comparing our approach with this baseline.
- **RGB+SDA**: SDA (semi-supervised discriminant analysis) [57] is a semi-supervised algorithm. Here it is applied to the RGB action videos in both the target and the auxiliary databases, which are considered as the sources of the labeled and the unlabeled training data in SDA, respectively. This baseline discards the depth maps and the skeleton structures captured by Kinect. It is used only in the experiments on IXMAS, i3DPost, and UIUC1 datasets, since the auxiliary dataset fully covers actions to be recognized in the three datasets.
- **Bor-DEP**: This baseline is a degenerate variant of our approach. Recall that our approach augments each RGB action video with additional depth maps and skeleton data. This baseline discards the original RGB features and the borrowed skeleton structures. It simply works on the borrowed depth maps. The adopted features for depth maps here are those described in Section V. Investigating the performance of this baseline helps identify whether the borrowed features themselves are informative or not.
- **Bor-SKE**: This baseline is the same as **Bor-DEP**, except the used data features are changed from the borrowed depth maps to the borrowed skeleton structures.
- **1NN-Bor**: For each action in the target database, we search its nearest neighbor in the auxiliary dataset according to their RGB features. The action is then augmented with the corresponding depth maps and skeleton structures. We also use the late fusion method for feature combination. This baseline doesn't take the inter-database variations into account, and directly borrows features without performing domain adaptation. By comparing with this baseline, the advantages of our approach, which jointly uses MRFs and KCCA to adapt and correlate the multi-modal features in different databases, can be revealed.
- **RGB+SKEGT**: In this baseline, actions come with the RGB features as well as the ground truth skeletons. Similar to our approach, late fusion is adopted to combine the two types of features. This baseline can be considered as the performance upper bound of our approach, since our approach augments actions with the estimated skeletons while the ground truth skeletons are given in this baseline. Note that this baseline is performed only in the experiments on UCF CIL dataset, in which the ground truth skeletons are provided.

In addition to these baselines, several published systems on the five benchmarks are also included for comparison.

C. Results in the Cases where Target Actions Are Fully Covered

We first evaluate our approach in IXMAS, i3DPost, and UIUC1 datasets, i.e., the cases where the auxiliary dataset covers all kinds of actions to be recognized. To make a fair comparison, we adopt the setup, Leave-One-Actor-Out (LOAO) cross validation, which is also used in [51], [52]. Suppose there are N actors in constructing a benchmark. LOAO in this case is the same as N-fold cross validation, except the training data by the same actor must belong to an identical fold. With LOAO, we in turn conduct the experiments on the three adopted benchmark datasets, and the auxiliary database we compiled is used in all the experiments. The recognition rates of our approach, the five baselines, and the state-of-the-art systems

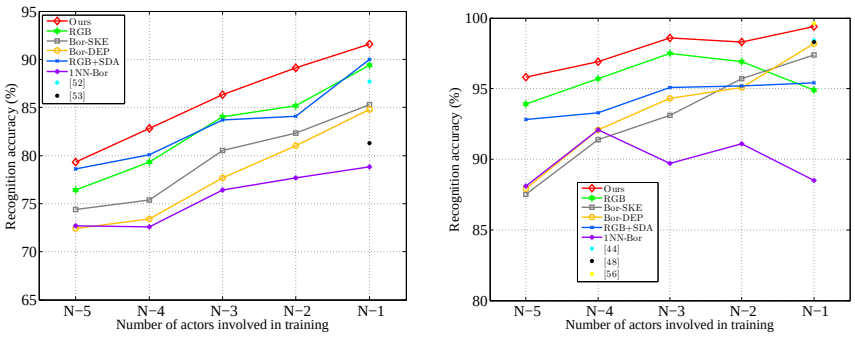


Fig. 5. The recognition rates of our approach and other baselines with different numbers of training actors in benchmarks (a) IXMAS where $N = 10$ and (b) UIUC1 where $N = 8$

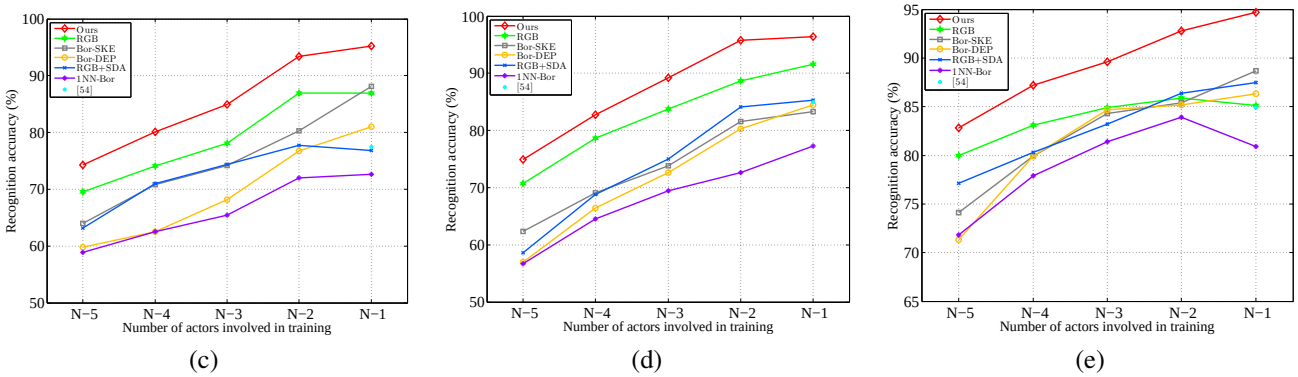


Fig. 6. The performances of various approaches with different numbers of training actors in benchmark i3Dpost where $N = 8$. The recognition rates in three different settings are plotted, including (a) single-view $0\,^\circ$, (b) single-view $45\,^\circ$, and (c) multi-view $0\,^\circ \cup 45\,^\circ$.

are reported in TABLE I, TABLE II, and TABLE III, one for each benchmark.

In TABLE I, the baseline RGB achieves recognition rate of 89.4% in IXMAS dataset, and it outperforms the state-of-the-art systems in the benchmark, such as [52], [53]. This is because we adopt the powerful descriptor [43] for feature extraction as well as the effective video processing tools [41], [42] for background removal. The baseline RGB-SDA is a bit better than baseline RGB, so the unlabeled RGB data in the auxiliary dataset contain useful information for regularizing the training of the classifier. The baseline 1NN-Bor cannot account for the inter-database variations. The borrowed features are very corrupt, and hence result in the performance degradation. One thing to be noted is that the baselines Bor-DEP and Bor-SKE get recognition rates of 85.3% and 84.8%, respectively. It points out that the borrowed features by the proposed mechanism are quite informative. Our approach further merges the three types of features, including the original RGB features, the borrowed depth and skeleton features, and results in a satisfactory performance, 91.6%.

As mentioned previously, we consider three different cases for performance evaluation in benchmark i3DPost, including two singe-view settings (single-view $0\,^\circ$ and single-view $45\,^\circ$ for abbreviation), and one multi-view setting (multi-view $0\,^\circ \cup 45\,^\circ$ for abbreviation). Thus, there are three sets of quantitative results shown in TABLE II, one for each case. The distribution of accuracy rates of various approaches in the benchmark is similar to that in IXMAS, but it is worth mentioning some interesting observations. The baseline Bor-DEP and Bor-SKE are comparable or even better than baseline RGB. This phenomenon indicates that depth maps and skeleton structures are discriminative for actions in i3DPost. Our approach can effectively borrow features across video modalities, and leverage both the original and the borrowed features to result in much better accuracy. The performance gains of our approach over baseline RGB are significant in all the three settings, i.e., 8.3% (95.2%-86.9%) in single-view $0\,^\circ$, 4.8% (96.4%-91.6%) in single-view $45\,^\circ$, and 9.6% (94.7%-85.1%) in multi-view $0\,^\circ \cup 45\,^\circ$. With the aid of cross-modal feature augmentation, our approach also remarkably outperforms the state-of-the-art system [54].

As can be seen in TABLE III, although the recent work [56] achieved the highest recognition accuracy in UIUC1, our approach is still competitive. It can boost the accuracy rate from 94.9% to 99.4%, an almost ideal performance. This is because our approach can successfully retrieve the corresponding depth maps and skeleton structures for each action to be recognized. The phenomenon will be further discussed later.

To demonstrate the power of the proposed strategy of information borrowing, we used small amount of training samples in the training stage to check if our approach can work well under poor training environments. Suppose again there are in total N actors in a dataset. We use the action data of $N-k$ actors for training, while the rest are used for testing. We

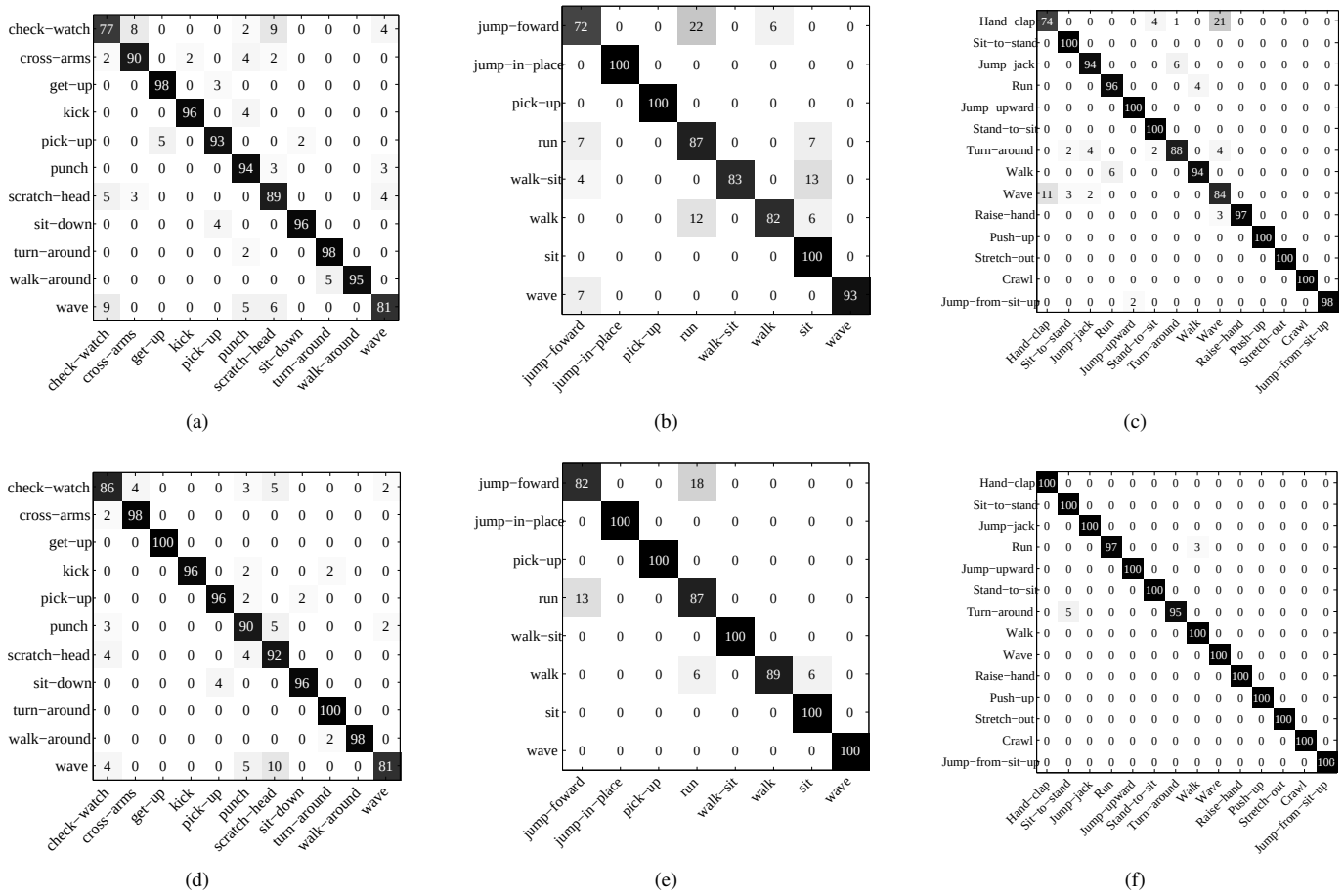


Fig. 7. The confusion tables (in %) by two approaches (baseline RGB and Ours) on the three benchmarks, IXMAS, i3DPost, and UIUC1. (a) Baseline RGB on IXMAS. (b) Baseline RGB on i3Dpost with the multi-view $0^\circ \cup 45^\circ$ setting. (c) Baseline RGB on UIUC1. (d) Ours on IXMAS. (e) Ours on i3Dpost with the multi-view $0^\circ \cup 45^\circ$ setting. (f) Ours on UIUC1.

respectively set $k = 1, 2, ..., 5$. In each case, we try 16 random splits of the actors, and compute the average recognition rates. The results by various approaches on IXMAS and UIUC1 are plotted in Fig. 5(a) and Fig. 5(b) respectively. Fig. 6 gives the outcomes with the three different camera settings in i3DPost dataset.

It can be observed in Fig. 5 and Fig. 6 that although the performances of all the approaches degenerate when the numbers of training samples decrease, our approach consistently achieves superior results to the five baselines and the other systems in most cases. The outcomes support that our approach can work robustly with different amounts of training data. Besides, the introduction of the auxiliary database indeed improves the performance and makes our approach outperform the state-of-the-art methods when only a small amount of training samples is provided. For instance in Fig. 6, our approach, in the case $N - 4$, still gives higher recognition rates than the state-of-the-art method [54] in the case $N - 1$. It demonstrates that our approach can make use of the auxiliary database to compensate for the lack of training data.

To gain insight into the quantitative results reported above, we investigate the performance improvement by our approach, especially from the viewpoint of feature expansion/augmentation. Recall that the main difference between our approach and the baseline RGB is that the augmented features are taken into account by the former, but ignored by the latter. Thus, we determine why augmented features help in this work by comparing the two approaches.

The confusion tables by baseline RGB and our approach on IXMAS dataset are given in Fig. 7(a) and in Fig. 7(d), respectively. One can observe that our approach gives much better accuracies than baseline RGB in class *check-watch* and *cross-arms*. It is evident that the performance gains in the two classes result from the introduction of the borrowed depth and skeleton features. On the other hand, both two approaches do not work well on class *wave*, relatively to other classes. The borrowed depth skeleton features do not help. We consider that the effectiveness of feature borrowing depends on whether the borrowed features are informative and complementary to the original features. The point of view is clarified through the examples of feature augmentation in Fig. 8. Note that the format of each of the twelve examples in Fig. 8 is the same as that in Fig. 3, except the action classes in ground truth (GT) and predicted by our approach and baseline RGB are also included.

To see why the borrowed features help recognize actions of class *check-watch*, an example of feature augmentation is given in Fig. 8(A1). Baseline RGB fails to correctly classify

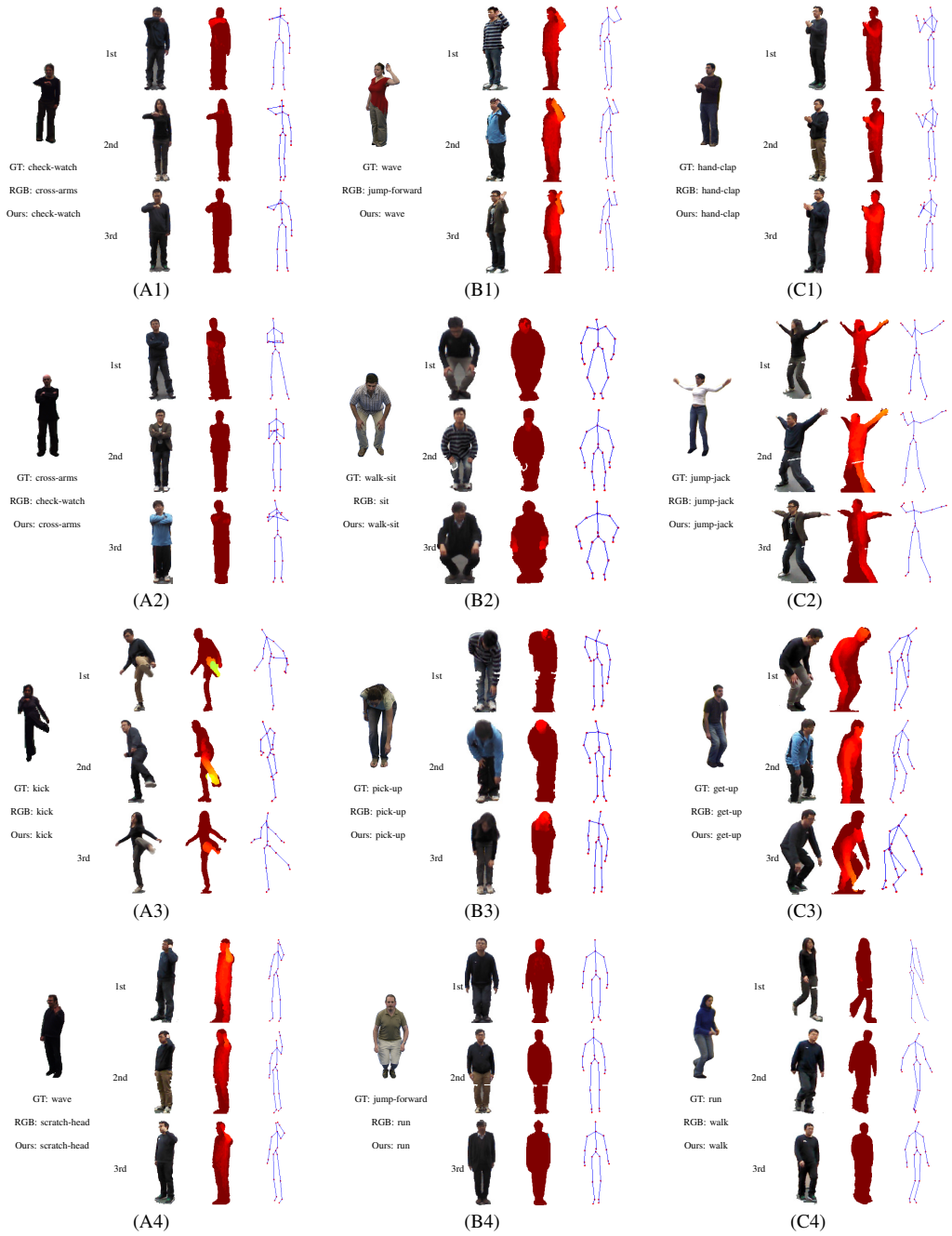


Fig. 8. Twelve examples of feature augmentation. The format of each example is the same as that in Fig. 3, except the action classes in ground truth (GT) and predicted by our approach and baseline RGB are also included. (A1) $\sim$ (A4) Four examples on IXMAS. (B1) $\sim$ (B4) Four examples on i3DPost with the multi-view $0^\circ \cup 45^\circ$ setting. (C1) $\sim$ (C4) Four examples on UIUC1.

this action in the example. Our approach instead exploits the borrowed features and makes the correct prediction. It can be checked that the borrowed depth maps and skeleton structures by our approach clearly highlight the forearms, the most discriminant part for distinguishing actions of this class from the rest. Similarly, an example for class *cross-arm* is given in Fig. 8(A2). Although feature borrowing improves the performance in most cases, we consider another example in Fig. 8(A4), in which both our approach and baseline RGB make wrong prediction. In this example, the highly similar appearances between actions of class *wave* and class *scratch-head* in side view make our approach fail to retrieve appropriate depth maps and skeleton structures.

The performance gains of our approach over baseline RGB on i3DPost are more remarkable than those on IXMAS. Fig. 7(b) and Fig. 7(e) respectively illustrate the confusion matrices by baseline RGB and our approach on i3DPost with the multi-view $0^\circ \cup 45^\circ$ setting. The accuracies in classes *jump-forward*, *walk-sit*, *walk*, and *wave* are significantly boosted by our approach. We also show some examples of feature

TABLE IV
RECOGNITION RATES (%) OF DIFFERENT METHODS ON THE UCF SPORTS DATASET.

Method	Ours	RGB	Bor-DEP	Bor-SKE	1NN-Bor	[13]	[44]
Accuracy (%)	94.0	86.0	63.3	64.7	87.3	88.2	89.1

TABLE V
RECOGNITION RATES (%) OF DIFFERENT METHODS ON THE UCF CIL DATASET.

Method	Ours	RGB	Bor-DEP	Bor-SKE	1NN-Bor	RGB+SKEGT	[58]	[59]
Accuracy (%)	98.3	81.0	70.7	77.6	82.8	100.0	95.8	100.0

augmentation in Fig. 8(B1) $\sim$ Fig. 8(B4). It can be visually and statistically checked that the borrowed depth and skeleton features are discriminant, and can complement the RGB features.

As for benchmark UIUC1, the confusion tables by the two approaches are plotted in Fig. 7(c) and Fig. 7(f). Although the RGB features have been good enough in this benchmark, our approach can still leverage the auxiliary database, and achieve nearly flawless recognition results. It can be seen in Fig. 8(C1) $\sim$ Fig. 8(C4) that the retrieved depth maps and skeleton structures quite match the given actions.

D. Results in the Cases where Target Actions Are Partially Covered

We have evaluated our approach in the cases where the actions to be recognized are fully covered by the auxiliary dataset. Then, more challenging cases are considered. Our approach is tested on the UCF sports and UCF CIL datasets. Large intra-class variations are involved in the two datasets. In addition, actions in the two datasets are only partially covered by the auxiliary dataset.

For comparing with existing methods, we follow the Leave-One-Out protocol used in [13], [44] for performance evaluation on UCF sports dataset, i.e., testing on each action while training on the rest together with their flipped versions. We follow the evaluation protocol suggested in [58], [59] to evaluate the performance on UCF CIL dataset. The recognition rates by our approaches, the adopted baselines, and the state-of-the-art methods on the two datasets are reported in TABLE IV and TABLE V, respectively.

As shown in TABLE IV, the proposed approach achieves recognition rate of 94.0%, and significantly outperforms baseline RGB. Our approach with the augmentation of the additional depth and skeleton features also performs better than the state-of-the-art systems [13], [44]. A major difference between testing on UCF sports dataset and on the first three datasets is that baselines Bor-DEP and Bor-SKE are no longer as powerful as baseline RGB. This is mainly caused by the fact that the action classes in UCF sports are not completely included in our auxiliary dataset. However, we also found that actions of different classes in UCF sports dataset tend to be associated with depth and skeleton features from actions of distinct classes in the auxiliary dataset. It results from the objective function of MRF in (3), in which discriminant learning is achieved by taking the labels of training data into account. It implies that the borrowed features can still enhance the performance of classification, even if they are not consistent with actions to be recognized in terms of action classes. It should be noted that there is a minor performance gap between baseline RGB and the approach in [44] though the same RGB features are used in both of them. We have tried to reproduce the approach in [44] as faithfully as possible. The minor performance gap may result from some tunable parameters used to convert videos from bags of features to histograms.

Like in the results of UCF sports dataset, similar performance rankings of our approach and the baselines can be found in the UCF CIL dataset. As shown in TABLE V, the recent work [59] achieved 100% recognition accuracy in UCF CIL dataset. Under the circumstance, our approach is still competitive. Our approach can boost the accuracy from 81.0% to 98.3% by borrowing and fusing the depth and skeleton features. It is also worth mentioning that baseline RGB+SKEGT serves as the performance upper bound of our approach since the ground truth of skeleton structures is used. It also achieves perfect performance in this dataset. However, there is only a minor performance drop, 1.7% $(= 100.0\% - 98.3\%)$, when the ground truth skeletons are replaced with the borrowed skeletons. It confirms that our approach can provide good alternatives to the unavailable skeletons of actions.

The confusion tables filled by the results of the two approaches, baseline RGB and Ours, on the UCF sports and UCF CIL datasets are plotted in Fig. 9. As can be seen, augmenting borrowed features improves the recognition rates of most actions categories in both datasets. We also visually compare the ground truth skeletons and the borrowed skeletons by our approach. Two examples of skeleton feature augmentation are shown in Fig. 10. Each example comes with an action in UCF CIL dataset, its ground truth skeleton structure, and the top three skeleton structures retrieved by our approach. The action categories of the ground truth skeletons and the borrowed skeletons are different, but their shapes are similar. Besides, the actions of different categories in UCF sports and UCF CIL datasets tend to be associated with skeleton features from actions of distinct classes in the auxiliary dataset. Thus, the borrowed skeletons are still helpful in improving action recognition.

In general, action videos in the five benchmarks often include self-occlusions or depth changes, especially in classes *cross-arm*, *punch*, and *walk-sit*. Compared with RGB features, depth maps are more discriminant in the cases. In addition, the appearance of an action changes dramatically with different

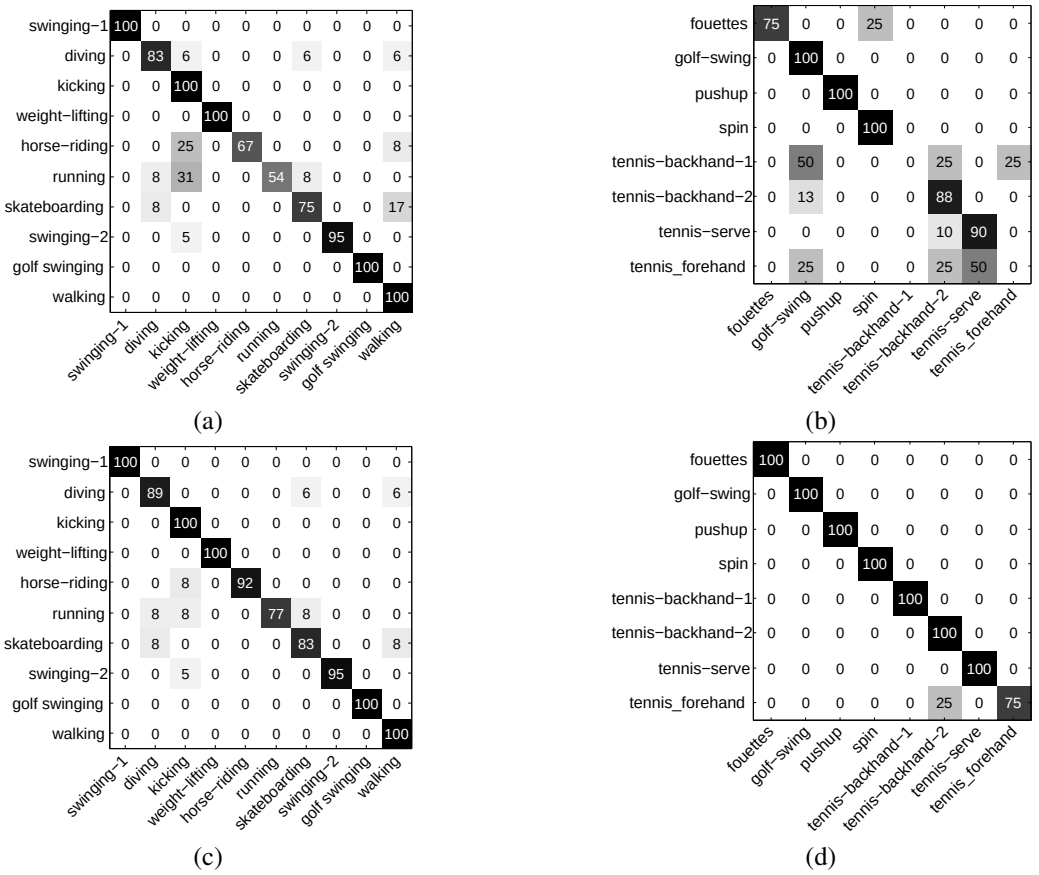


Fig. 9. The confusion tables (in %) by two approaches (baseline RGB and Ours) on the benchmarks, UCF sports and UCF CIL. (a) Baseline RGB on UCF sports. (b) Baseline RGB on UCF CIL. (c) Ours on UCF sports. (d) Ours on UCF CIL.

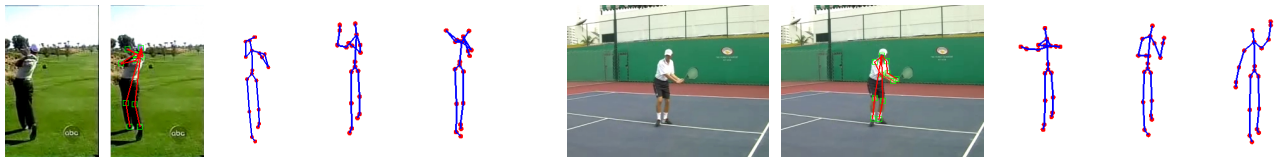


Fig. 10. Two examples of skeleton feature augmentation. (a) & (f) Two actions in UCF CIL dataset. (b) & (g) The corresponding ground truth skeleton structures. (c) $\sim$ (e) & (h) $\sim$ (j) the top three skeleton structures retrieved by our approach.

angles of view. The 3D skeleton structures in the situations are more robust to this type of intra-class variations. Our approach can effectively retrieve depth and skeleton features for actions to be recognized. It hence facilitates the accomplishment of a more accurate action recognition system.

VII. Conclusions

The new types of imaging devices provide the opportunity of better solving increasingly complex video processing tasks, but their respective limitations are currently hindering the practical applicability. In the work, we resolve this problem by proposing an approach that can borrow information from an offline collected database where multi-modal videos taken by heterogeneous cameras are available. Promising experimental results demonstrate that our approach can effectively adapt the variations between different databases, transfer knowledge across video modalities, and lead to remarkable performance boosting. Our approach hence provides an alternative way of utilizing the emerging cameras even in the cases where they are not online accessible. In addition, the proposed approach is developed to carry out cross-modal information borrowing in a general way. It can be applied to a set of applications where multiple video modalities are appreciated, such as gesture recognition, human pose estimation, scene understanding, content-based image/video analysis and retrieval.

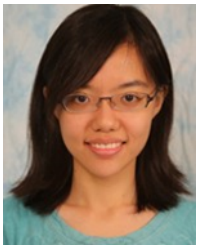
References

[1] R. Poppe, "A survey on vision-based human action recognition," *Image and Vision Computing*, vol. 28, no. 6, pp. 976–990, 2010.
[2] "Microsoft Kinect," http://en.wikipedia.org/wiki/Kinect.
[3] "FUJIFILM FinePix 3D," http://en.wikipedia.org/wiki/Fujifilm_FinePix_Real_3D.
[4] "FLIR T620," http://www.flir.com.
[5] "Lytro," http://en.wikipedia.org/wiki/Lytro.
[6] 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 *Proc. Conf. Computer Vision and Pattern Recognition*, 2011, pp. 1297–1304.
[7] W. Shen, K. Deng, X. Bai, T. Leyvand, B. Guo, and Z. Tu, "Exemplar-based human action pose correction and tagging," in *Proc. Conf. Computer Vision and Pattern Recognition*, 2012, pp. 1784–1791.
[8] S. Li, *Markov Random Field Modeling in Image Analysis*. Springer, 2009.
[9] F. Bach and M. Jordan, "Kernel independent component analysis," *J. Machine Learning Research*, vol. 3, pp. 1–48, 2002.
[10] A. Bobick and J. Davis, "The recognition of human movement using temporal templates," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 23, no. 3, pp. 257–267, 2001.
[11] L. Gorelick, M. Blank, E. Shechtman, M. Irani, and R. Basri, "Actions as space-time shapes," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 29, no. 12, pp. 2247–2253, 2007.
[12] I. Laptev and T. Linderberg, "Space-time interest points," in *Proc. Int'l Conf. Computer Vision*, 2003, pp. 432–439.
[13] H. Wang, A. Kläser, C. Schmid, and C. Liu, "Action recognition by dense trajectories," in *Proc. Conf. Computer Vision and Pattern Recognition*, 2011, pp. 3169–3176.
[14] P. Matikainen, M. Hebert, and R. Sukthankar, "Representing pairwise spatial and temporal relations for action recognition," in *Proc. Euro. Conf. Computer Vision*, 2010, pp. 508–521.
[15] K. Prabhaka, S. Oh, P. Wang, G. D. Abowd, and J. M. Rehg, "Temporal causality for the analysis of visual events," in *Proc. Conf. Computer Vision and Pattern Recognition*, 2010, pp. 1967–1974.
[16] L. Wang and D. Suter, "Recognizing human activities from silhouettes: motion subspace and factorial discriminative graphical model," in *Proc. Conf. Computer Vision and Pattern Recognition*, 2007.
[17] C. Chen and J. Aggarwal, "Modeling human activities as speech," in *Proc. Conf. Computer Vision and Pattern Recognition*, 2011, pp. 3425–3431.
[18] R. Girshick, J. Shotton, P. Kohli, A. Criminisi, and A. Fitzgibbon, "Efficient regression of general-activity human poses from depth images," in *Proc. Int'l Conf. Computer Vision*, 2011, pp. 415–422.
[19] X. Yang, C. Zhang, and Y. Tian, "Recognizing actions using depth motion maps-based histograms of oriented gradients," in *Proc. ACM Conf. Multimedia*, 2012, pp. 1057–1060.
[20] "OpenNI Library," http://www.openni.org/.
[21] J. Wang, Z. Liu, Y. Wu, and J. Yuan, "Mining actionlet ensemble for action recognition with depth cameras," in *Proc. Conf. Computer Vision and Pattern Recognition*, 2012, pp. 1290–1297.
[22] C. Ellis, S. Z. Masood, M. F. Tappen, J. J. LaViola, and R. Sukthankar, "Exploring the trade-off between accuracy and observational latency in action recognition," *Int. J. Computer Vision*, vol. 101, no. 3, pp. 420–436, 2013.
[23] N. Ashraf, C. Sun, and H. Foroosh, "View invariant action recognition using projective depth," *Computer Vision and Image Understanding*, vol. 123, pp. 41–52, 2014.
[24] N. C. Tang, Y.-Y. Lin, J.-H. Hua, M.-F. Weng, and H.-Y. M. Liao, "Human action recognition using associated depth and skeleton information," in *Proc. Int'l Conf. Acoustics, Speech, and Signal Processing*, 2014, pp. 4608–4612.
[25] S. Pan and Q. Yang, "A survey on transfer learning," *IEEE Trans. Knowledge and Data Engineering*, vol. 22, no. 10, pp. 1345–1359, 2010.
[26] L. Jie, T. Tommasi, and B. Caputo, "Multiclass transfer learning from unconstrained priors," in *Proc. Int'l Conf. Computer Vision*, 2011, pp. 1863–1870.
[27] L. Cao, Z. Liu, and T. Huang, "Cross-dataset action detection," in *Proc. Conf. Computer Vision and Pattern Recognition*, 2010, pp. 1998–2005.
[28] Q. Yin, X. Tang, and J. Sun, "An associate-predict model for face recognition," in *Proc. Conf. Computer Vision and Pattern Recognition*, 2011, pp. 497–504.
[29] L. Torresani, M. Szummer, and A. Fitzgibbon, "Efficient object category recognition using classemes," in *Proc. Euro. Conf. Computer Vision*, 2010, pp. 776–789.
[30] M.-F. Weng and Y.-Y. Chuang, "Cross-domain multi-cue fusion for concept-based video indexing," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 34, no. 10, pp. 1927–1941, 2012.
[31] C. Snoek, M. Worring, and A. Smeulders, "Early versus late fusion in semantic video analysis," in *Proc. ACM Conf. Multimedia*, 2005, pp. 399–402.
[32] G. Lanckriet, N. Cristianini, P. Bartlett, L. Ghaoui, and M. Jordan, "Learning the kernel matrix with semidefinite programming," *J. Machine Learning Research*, vol. 5, pp. 27–72, 2004.
[33] A. Rakotomamonjy, F. Bach, S. Canu, and Y. Grandvalet, "SimpleMKL," *J. Machine Learning Research*, vol. 9, p. 2491V2521, 2008.
[34] S. Sonnenburg, G. Rätsch, C. Schäfer, and B. Schölkopf, "Large scale multiple kernel learning," *J. Machine Learning Research*, vol. 7, pp. 1531–1565, 2006.
[35] P. Gehler and S. Nowozin, "On feature combination for multiclass object classification," in *Proc. Int'l Conf. Computer Vision*, 2009, pp. 221–228.
[36] Y.-Y. Lin, T.-L. Liu, and C.-S. Fuh, "Multiple kernel learning for dimensionality reduction," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 33, no. 6, pp. 1147–1160, 2011.
[37] Y. Boykov, O. Veksler, and R. Zabih, "Fast approximate energy minimization via graph cuts," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 23, no. 11, pp. 1222–1239, 2001.
[38] B. Schölkopf, A. Smola, and K.-R. Müller, "Nonlinear component analysis as a kernel eigenvalue problem," *Neural Computation*, vol. 10, no. 5, pp. 1299–1319, 1998.
[39] B. Schölkopf and A. Smola, *Learning with Kernels*. MIT Press, 2002.
[40] S. Roweis and L. Saul, "Nonlinear dimensionality reduction by locally linear embedding," *Science*, vol. 290, pp. 2323–2326, 2000.
[41] 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," *IEEE Trans. Multimedia*, vol. 13, no. 4, pp. 602–614, 2011.
[42] O. Barnich and M. V. Droogenbroeck, "ViBe: A universal background subtraction algorithm for video sequences," *IEEE Trans. Image Processing*, vol. 20, no. 6, pp. 1709–1724, 2011.
[43] D. Weinland, M. Özuysal, and P. Fua, "Making action recognition robust to occlusions and viewpoint changes," in *Proc. Euro. Conf. Computer Vision*, 2010, pp. 635–648.
[44] H. Wang, A. Kläser, C. Schmid, and C.-L. Liu, "Dense trajectories and motion boundary descriptors for action recognition," *Int. J. Computer Vision*, vol. 103, no. 1, pp. 60–79, 2013.
[45] G. Zhao and M. Pietikainen, "Dynamic texture recognition using local binary patterns with an application to facial expressions," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 29, no. 6, pp. 915–928, 2007.
[46] D. Weinland, E. Boyer, and R. Ronfard, "Action recognition from arbitrary views using 3D exemplars," in *Proc. Int'l Conf. Computer Vision*, 2007.
[47] 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*, 2009, pp. 159–168.
[48] D. Tran and A. Sorokin, "Human activity recognition with metric learning," in *Proc. Euro. Conf. Computer Vision*, 2008, pp. 548–561.
[49] M. Rodriguez, J. Ahmed, and M. Shah, "Action MACH a spatio-temporal maximum average correlation height filter for action recognition," in *Proc. Conf. Computer Vision and Pattern Recognition*, 2008.
[50] Y. Shen and H. Foroosh, "View-invariant recognition of body pose from space-time templates," in *Proc. Conf. Computer Vision and Pattern Recognition*, 2008.
[51] G. Burghouts and K. Schutte, "Spatio-temporal layout of human actions for improved bag-of-words action detection," *Pattern Recognition Letters*, vol. 34, no. 15, pp. 1861–1869, 2013.
[52] X. Wu, D. Xu, L. Duan, and J. Luo, "Action recognition using context and appearance distribution features," in *Proc. Conf. Computer Vision and Pattern Recognition*, 2011, pp. 489–496.
[53] M. Varma and B. R. Babu, "More generality in efficient multiple kernel learning," in *Proc. Int'l Conf. Machine Learning*, 2009.
[54] A. Iosifidis, A. Tefas, and I. Pitas, "View-invariant action recognition based on artificial neural networks," *IEEE Trans. Neural Networks and Learning Systems*, vol. 23, no. 3, pp. 412–424, 2012.
[55] A. Iosifidis, Tefas, and I. Pitas, "Multi-view action recognition based on action volumes fuzzy distances and cluster discriminant analysis," *Signal Processing*, vol. 93, no. 6, pp. 1445–1457, 2013.
[56] J. Hernandez, R. Cabido, A. S. Montemayor, and J. Pantrigo, "Human activity recognition based on kinematic features," *Expert Systems*, 2013.
[57] D. Cai, X. He, and J. Han, "Semi-supervised discriminant analysis," in *Proc. Int'l Conf. Computer Vision*, 2007.
[58] Y. Shen and H. Foroosh, "View-invariant action recognition using fundamental ratios," in *Proc. Conf. Computer Vision and Pattern Recognition*, 2008.
[59] Y.-P. Shen and H. Foroosh, "View-invariant action recognition from point triplets," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 31, no. 10, pp. 1898–1905, 2009.

Nick C. Tang received the B.S. and M.S. degrees from Tamkang University, Tamsui, Taiwan, in 2003 and 2005, respectively. He also received the Ph.D. degree from Tamkang University in 2008. Currently, he is a Postdoctoral Fellow with the Institute of Information Science, Academia Sinica, Taipei, Taiwan. His research interests include image and video analysis, computer vision, computer graphics, and their applications.

Yen-Yu Lin received the B.S. degree in information management, and the M.S. and Ph.D. degrees in computer science and information engineering from National Taiwan University, Taipei, Taiwan, in 2001, 2003, and 2010, respectively. He is currently an Assistant Research Fellow with the Research Center for Information Technology Innovation, Academia Sinica, Taipei, Taiwan. His current research interests include computer vision, pattern recognition, and machine learning. He is a member of the IEEE.



Ju-Hsuan Hua received her B.S. in electrical engineering from National Taiwan University, Taipei, Taiwan, and her M.S. in robotics from Carnegie Mellon University, Pittsburgh, USA. She was involved in this research during her work as a research assistant in Academia Sinica, Taipei, Taiwan from 2012 to 2013.

Shih-En Wei received the B.S. degree in electrical engineering and M.S. degree in communication engineering from National Taiwan University, Taiwan, in 2010 and 2012, respectively. In 2013 to 2014, he was a full-time research assistant in the Institute of Information Science, Academia Sinica, Taiwan. Currently, he is a masters student in the Robotics Institute, Carnegie Mellon University, PA. His research interests include computer vision and machine learning.

Ming-Fang Weng received the B.S. degree and M.S. degree from National Chiao Tung University, Hsinchu, Taiwan, in 1998 and 2000 respectively, Ph.D. degree from National Taiwan University, Taipei, Taiwan, in 2010, all in computer science and information engineering. He was a Postdoctoral Fellow in the Institute of Information Science, Academia Sinica, Taipei, Taiwan, and is currently a Principal Engineer in the Institute for Information Industry, Taipei, Taiwan. His research interests include digital content analysis, image/video information retrieval, computer vision, and multimedia applications.

Hong-Yuan Mark Liao received his Ph.D. degree in electrical engineering from Northwestern University, Evanston, IL, USA, in 1990. In 1991, he joined the Institute of Information Science, Academia Sinica, Taipei, Taiwan, and is currently a Distinguished Research Fellow. He has worked in the fields of multimedia signal processing, image processing, computer vision, pattern recognition, video forensics, and multimedia protection for more than 25 years. During 2009-2011, he was the Division Chair of the Computer Science and Information Engineering Division II, National Science Council of Taiwan. He is jointly appointed as a Professor with the Department of Computer Science and Information Engineering, National Chiao Tung University, Hsinchu, Taiwan, and the Department of Electrical Engineering and Computer Science of National Cheng Kung University, Tainan, Taiwan. From 2009 to 2012, he was jointly appointed as the Multimedia Information Chair Professor of National Chung Hsing University, Taichung, Taiwan. Since 2010, he has been appointed as an Adjunct Chair Professor of Chung Yuan Christian University, Zhongli, Taiwan. Since 2014, he has been appointed as the Honorary Chair Professor of National Sun Yat-sen University, Kaohsiung, Taiwan. He was a recipient of the Young Investigators' Award from Academia Sinica in 1998, the Distinguished Research Award from the National Science Council of Taiwan in 2003, 2010, and 2013, the National Invention Award of Taiwan in 2004, the Distinguished Scholar Research Project Award from National Science Council of Taiwan in 2008, and the Academia Sinica Investigator Award in 2010. His professional activities include, the Co-Chair of the 2004 International Conference on Multimedia and Exposition (ICME), the Technical Co-Chair of 2007 ICME, the General Co-Chair of the 17th International Conference on Multimedia Modeling, the President of the Image Processing and Pattern Recognition Society of Taiwan (2006–2008), an Editorial Board Member of the *IEEE Signal Processing Magazine* (2010–2013), and an Associate Editor of the *IEEE Transactions on Image Processing* (2009–2013), the *IEEE Transactions on Information Forensics and Security* (2009–2012), and the *IEEE Transactions on Multimedia* (1998–2001). He has been a Fellow of the IEEE since 2013 for contributions to image and video forensics and security. He also serves as the IEEE Signal Processing Society Region 10 Director (Asia-Pacific Region).