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], and Hong-Yuan Mark Liao[1]

[1]Academia Sinica, Taiwan; [2]Carnegie Mellon University, USA

1. Summary

- We proposed a method to deal with the short effective distance problem of RGB-D cameras to take advantage of RGB-D cameras which are not online accessible. We illustrated it with the application to action recognition.
- Our approach
  - We use Kinect to offline collect a multi-modal database.
  - Our goal is to augment actions to be recognized with proper depth and skeleton features.
  - We do domain adaptation (DA), by which the actions to be recognized can be concisely reconstructed by entries in the auxiliary database.
  - Each action can be augmented with additional depth and skeleton images retrieved from the auxiliary database, and we use multiple kernel learning (MKL) to fuse these three kinds of features for action recognition.

2. Related Work

2.1 Feature Descriptor

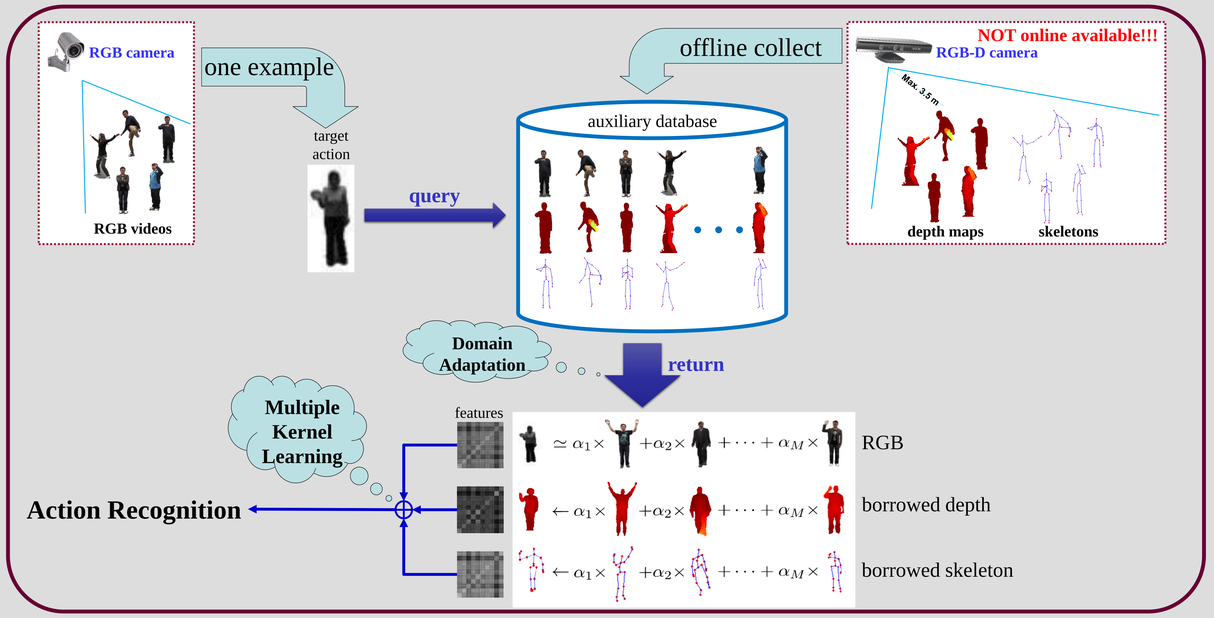
- Designing powerful feature descriptors against intra-class variations has gained significant progress.
  - Global: sensitive to occlusion and deformation
  - Local: less discriminant power
- Descriptors based on RGB data are vulnerable to camera perspectives and partial occlusions.
- RGB-D cameras can provide depth data, which affords informative clues for action recognition.
- *OpenNI library* provides the positions of key joints on the human body, *i.e.*, the skeleton, which is also helpful for action recognition.
- The short ranges of effective distances still make RGB-D cameras in real-world applications, e.g., surveillance.

2.2 Transfer Learning

- Training data acquisition is also a challenge.
  - Complete training data may not be available.
  - Labeling huge data is expensive.
- Transfer learning can alleviate the above problems.
- Our approach also utilizes knowledge transferred from an auxiliary database via DA and data reconstruction for improving action recognition.
- Difference: we borrow visual features across different video modalities, and resolves the problems caused by the absence of RGB-D cameras.

2.3 Multiple Kernel Learning

- MKL derives an optimal kernel over a given convex set of kernels, which means it can fuse heterogeneous features and lead to boosted performance.
- In our case, we fuse the original RGB features as well as the augmented depth and skeleton features.



3. The Proposed Method

- Goal: borrow depth and skeleton information from an auxiliary, multi-modal database.
- Symbols:
  - Target actions: $D = \{\boldsymbol{x}_i, y_i\}_{i=1}^{N}$, $\boldsymbol{x}_i$: RGB data, y_i: label $\in \{1,2,\dots,C\}$
  - Auxiliary database: $\widetilde{D} = \{\widetilde{\boldsymbol{x}}_i, \widetilde{\boldsymbol{d}}_i, \widetilde{\boldsymbol{s}}_i\}_{i=1}^{M}$, $\widetilde{\boldsymbol{d}}_i$: depth data, $\widetilde{\boldsymbol{s}}_i$: skeleton data
- We focus on associating each action $x_i \in D$ with proper depth map d_i and skeleton structure s_i, so that the absence of RGB-D cameras is compensated.

3.1 Domain Adaptation

- We adopt reconstruction-based DA to tackle the inter-database variations, which means we model D and $\widetilde{D}$ by a linear transformation and reconstruction.

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

- With discriminant learning and outlier handling, we obtain

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

- For optimization, we adopt some modifications
  - Serve nuclear norm as a convex approximation of rank minimization
  - Introduce an auxiliary variable $F = [F^1 \cdots F^c]$
  - Orthogonalize W afterwards with orthogonality preserving method

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

3.2 Optimization

- We optimize (3) by the inexact Augmented Lagrange Multiplier (ALM) method, which minimizes the augmented Lagrange function of (3):

$$\min_{W,A,F,E,U,V} \sum_{c=1}^{C} \|F^c\|_* + \lambda\|E\|_{2,1} + \langle U, A - F\rangle + \frac{\mu}{2}\|A - F\|_F^2 + \langle V, WX - \tilde{X}A - E\rangle + \frac{\mu}{2}\|WX - \tilde{X}A - E\|_F^2 \tag{4}$$

- Alternate optimization for $\{W, A, F, E\}$
  - 1. F: use singular value shrinkage operator
  - 2. W: closed-form solution
  - 3. Apply QR-decomposition to orthogonalize W
  - 4. E: use the analytical solution
  - 5. A: closed-form solution
  - 6. update Lagrange multipliers and penalty parameters
  - 7. check convergence conditions

3.3 Feature Augmentation

- After obtaining transformation W, we optimize reconstruction coefficients for both training and unseen testing data by solving

$$\boldsymbol{\alpha} = arg\min_{\boldsymbol{\alpha}} \left\|W\boldsymbol{x} - \tilde{X}\boldsymbol{\alpha}\right\|^2 + \gamma\|\boldsymbol{\alpha}\|^2 \tag{5}$$

- Use the obtained $\boldsymbol{\alpha}$ to reconstruct depth and skeleton features:

$$\boldsymbol{d} \leftarrow [\widetilde{\boldsymbol{d}}_1 \cdots \widetilde{\boldsymbol{d}}_M]\boldsymbol{\alpha} \ \text{ and } \ \boldsymbol{s} \leftarrow [\widetilde{\boldsymbol{s}}_1 \cdots \widetilde{\boldsymbol{s}}_M]\boldsymbol{\alpha} \tag{6}$$

- We compile an kernel matrix for actions in each modality, and adopt simpleMKL [Rakotomamonjy et al. *JMLR* 2008] to learn classifiers that optimally combine the three types of features.

4. Experimental Results

4.1 Data and Setting

- Three benchmarks are used for evaluation
  - IXMAS, i3DPost, UIUC-1: daily activities, multiple view points
- We use Kinect to build up a three-modal auxiliary database
  - 10 actors, 40 actions, two view angles

- Feature representations
  - RGB: 3D-HOG descriptor [Weinland et al. *ECCV* 2010]
  - Depth: spatio-temporal local binary pattern [Zhao et al. *TPAMI* 2007]
  - Skeleton: Fourier temporal pyramid [Wang et al. *CVPR* 2012]
- Two experiment settings
  - Leave-one-actor-out (LOAO): N-fold cross validation
  - Different numbers of actors: k for training, the rest for testing (k = 3 ~ N-1)

4.2 Baselines

- Goal: identify the contributions of the components in our approach
- RGB: ignore auxiliary database
- Bor-DEP: only use the associated depth maps
- Bor-SKE: only use the associated skeleton structures
- KSDA: consider RGB videos in D and $\widetilde{D}$ as the labeled and unlabeled training data in kernel SDA [Cai et al. *ICCV* 2007]
- 1NN-Bor: associate the nearest sample in $\widetilde{D}$ and borrow the corresponding depth and skeleton features (MKL is used.)
- Our approach
  - d+s: RGB + associated depth and skeleton
  - d: RGB + associated depth
  - s: RGB + associated skeleton

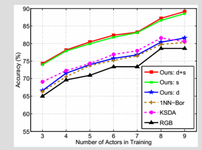
4.3 Results

- LOAO ([11], [12], [26], [32] are state-of-the-art methods)
  - IXMAS
  - i3DPost
  - UIUC-1

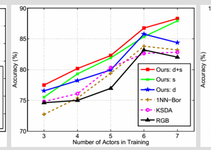
Method	Ours: d+s	Ours: d	Ours: s	RGB	Bor-DEP	Bor-SKE	KSDA	1NN-Bor	[32]
Accuracy	**89.1**	81.6	88.5	78.6	51.2	82.6	80.6	80.3	87.7
Method	Ours: d+s	Ours: d	Ours: s	RGB	Bor-DEP	Bor-SKE	KSDA	1NN-Bor	[12]
Accuracy	**88.3**	84.4	87.9	82.0	57.8	80.1	82.8	83.2	84.9
Method	Ours: d+s	Ours: d	Ours: s	RGB	Bor-DEP	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**

- Different numbers of actors

IXMAS



i3DPost



UIUC-1