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

To Appear in CVPR'14


Y.-Y. Lin


J.-H. Hua

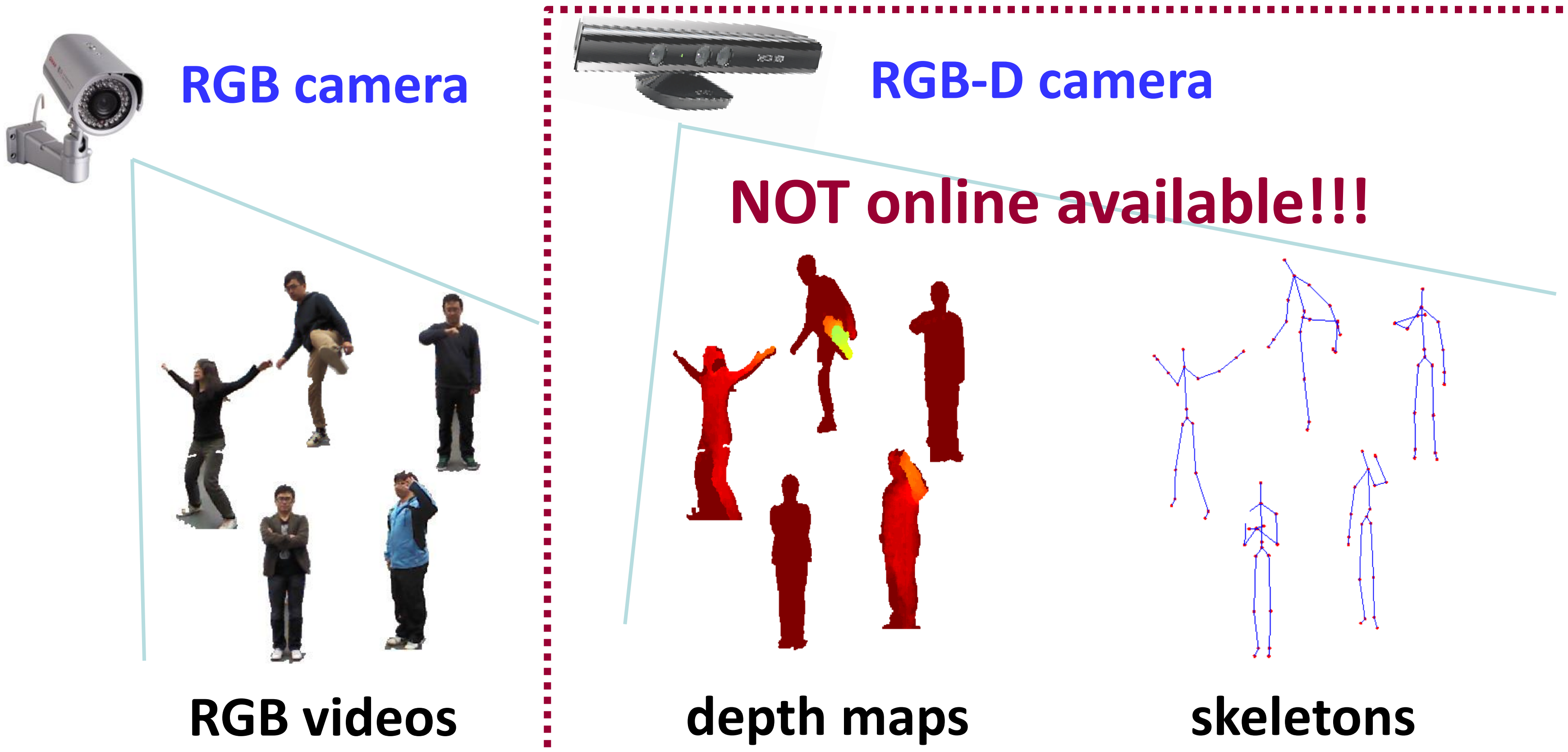

N. C. Tang


M.-H. Chen


H.-Y. M. Liao

The goal

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



Computer vision with next-generation cameras

- Computer vision
  - Let computers see, recognize, and interpret the world like humans
- CV techniques are highly adapted to imaging devices
  - Most existing techniques are developed on RGB images
- Recent advances in imaging devices

RGB-D

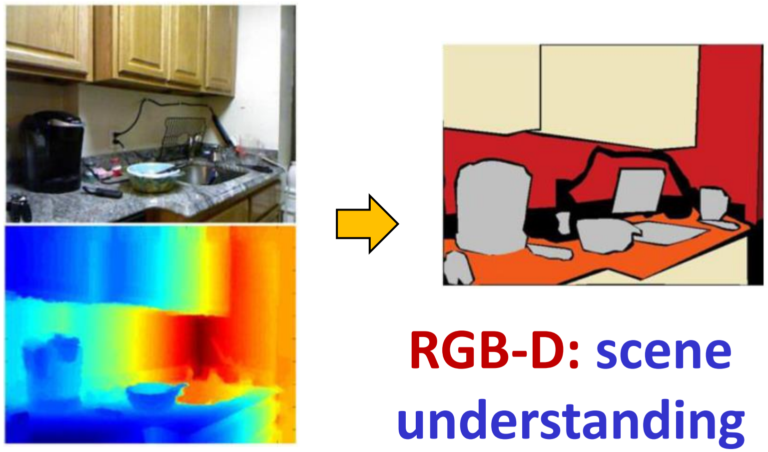
Binocular

High-speed

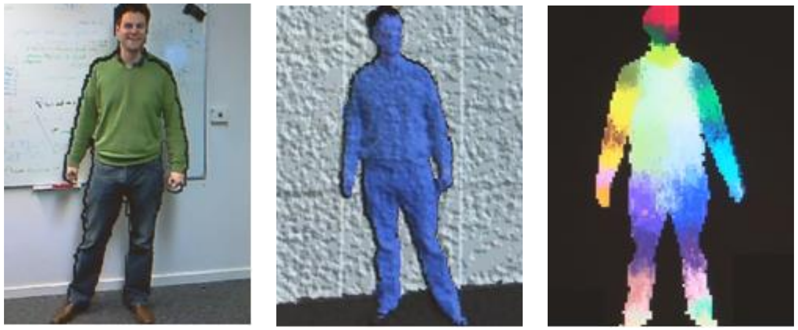
Lightfield

Infrared

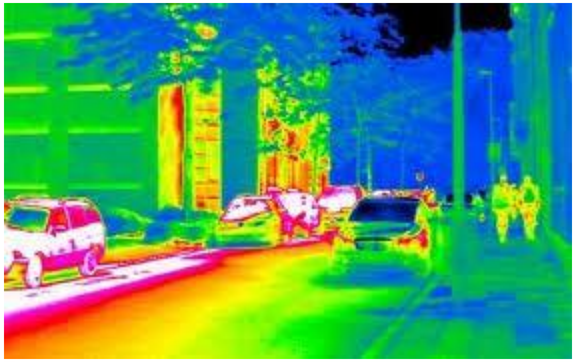
Their applications



RGB-D: scene understanding



RGB-D: pose estimation & action recognition



Infrared: night vision

Binocular: stereo vision

Research directions with emerging cameras

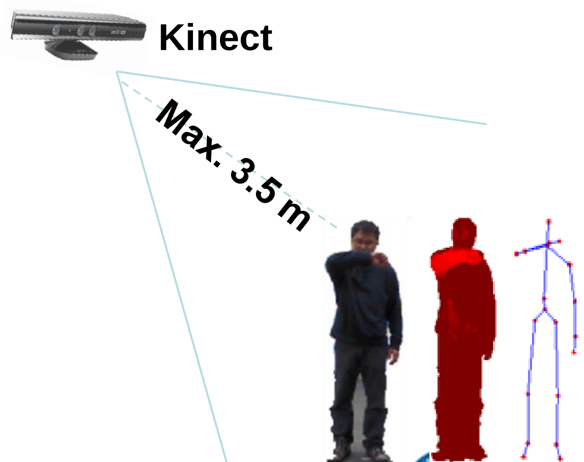
- Design new image descriptors and feature extractors
- Develop new machine learning algorithms
- Initiate new computer vision applications
- Address the limitations of these emerging cameras
  - Short range of the effective distance
  - Expensive cost
  - Long image processing time

Research directions with emerging cameras

- Design new image descriptors and feature extractors
- Develop new machine learning algorithms
- Initiate new computer vision applications
- Address the limitations of these emerging cameras
  - Short range of the effective distance in RGB-D cameras
  - Expensive cost
  - Long image processing time

The problem

- RGB-D cameras better solve **CV** applications
  - Scene understanding, action recognition, post estimation, object segmentation, …
- Microsoft Kinect: one of the most popular RGB-D cameras
  - Helpful for action recognition
  - Short effective distance: 1.2 ~ 3.5 meters
- The problem: Less applicability
  - Kinect is not online accessible in many real-world applications, e.g., surveillance

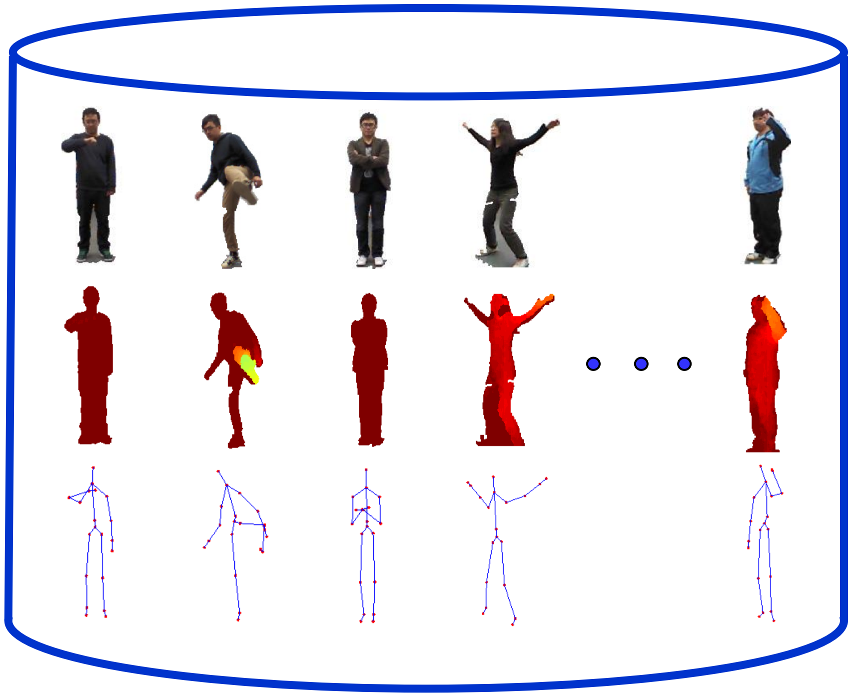


Our idea

- Propose an alternative scenario to address this problem, and illustrate it with the application to action recognition
- In most cases, we focus on recognizing predefined classes of actions in most applications
- Offline collect an auxiliary, multi-modal database by Kinect
  - Unsupervised
  - At least cover actions of interest
  - RGB videos, depth maps, and skeleton structures
- Depth-associated action recognition with the aid of the auxiliary database

Our idea

- Three-modal auxiliary database

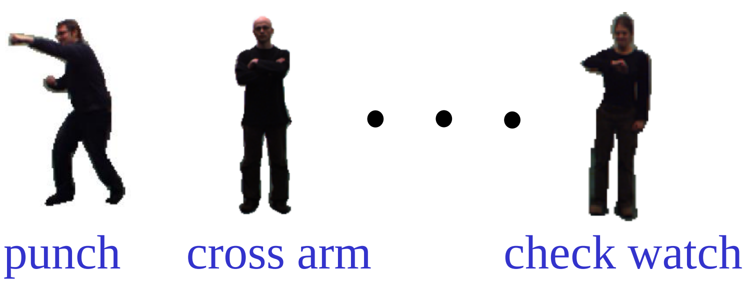


- Can the auxiliary database be an alternative to Kinect, and how?

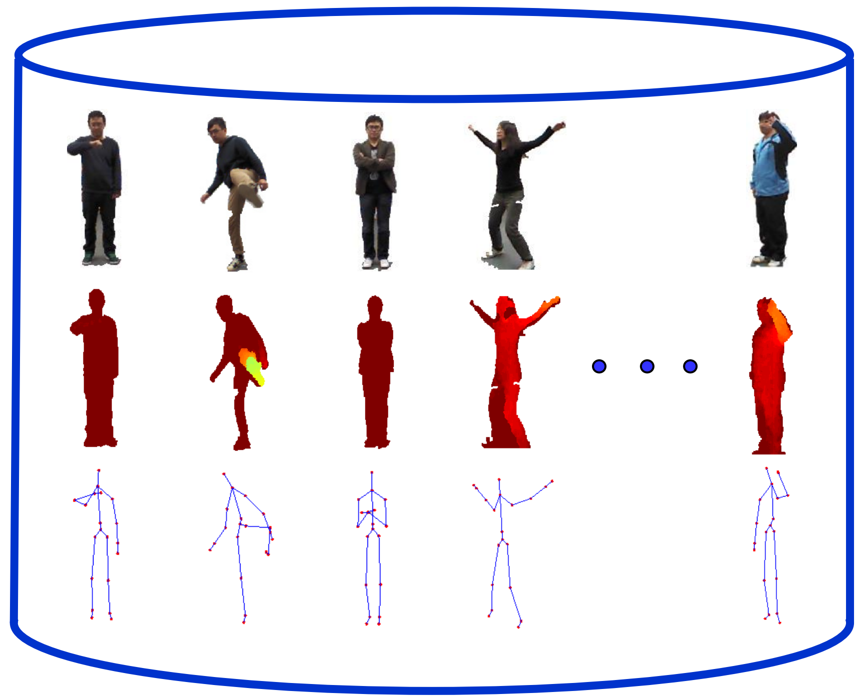
Action Recognition with An Auxiliary Database

- Action recognition as a multi-class classification problem
- RGB-D camera helps, but suffers from the short effective distance
- How to improve the performance if an auxiliary, multi-modal database is available

Training Phase:

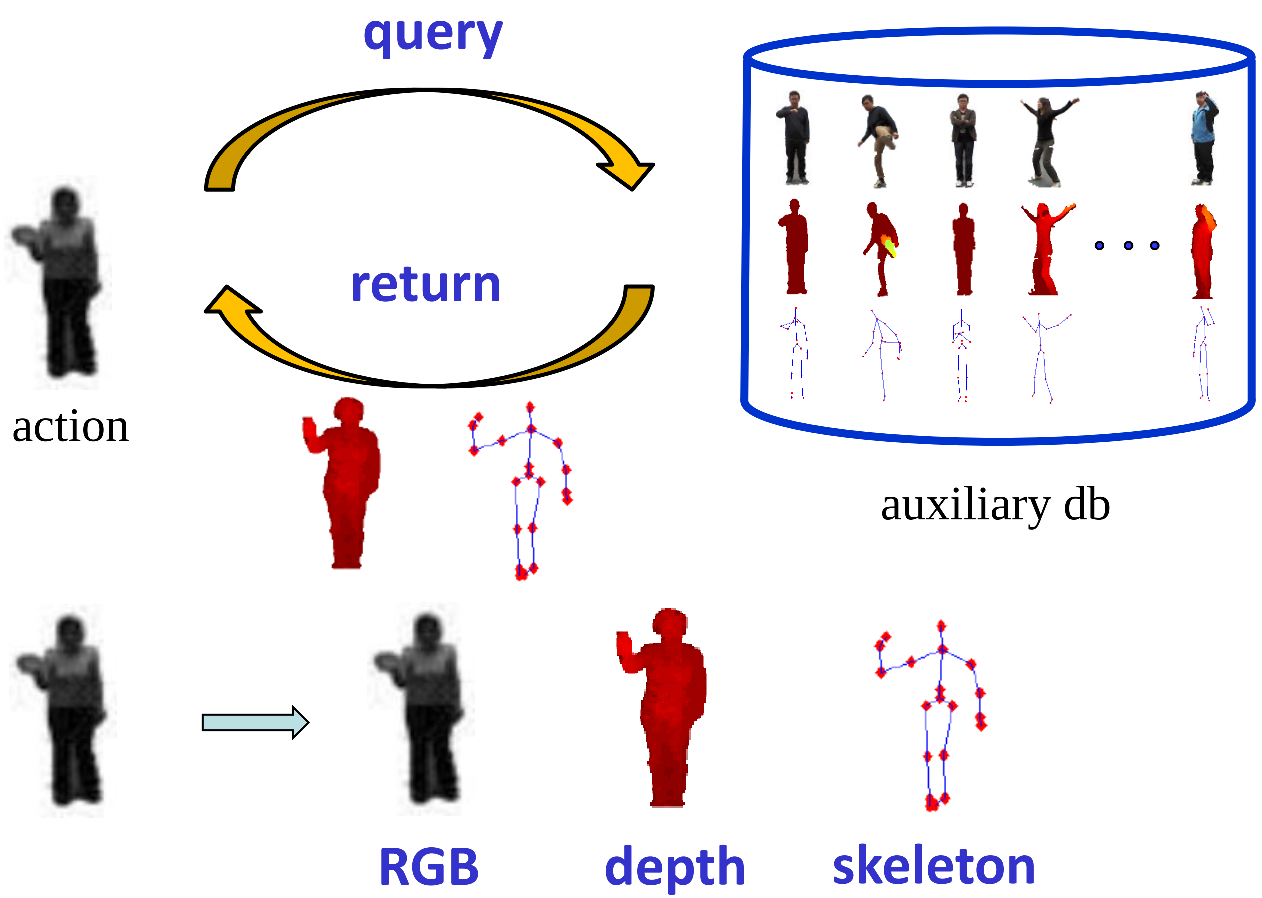


Testing Phase:



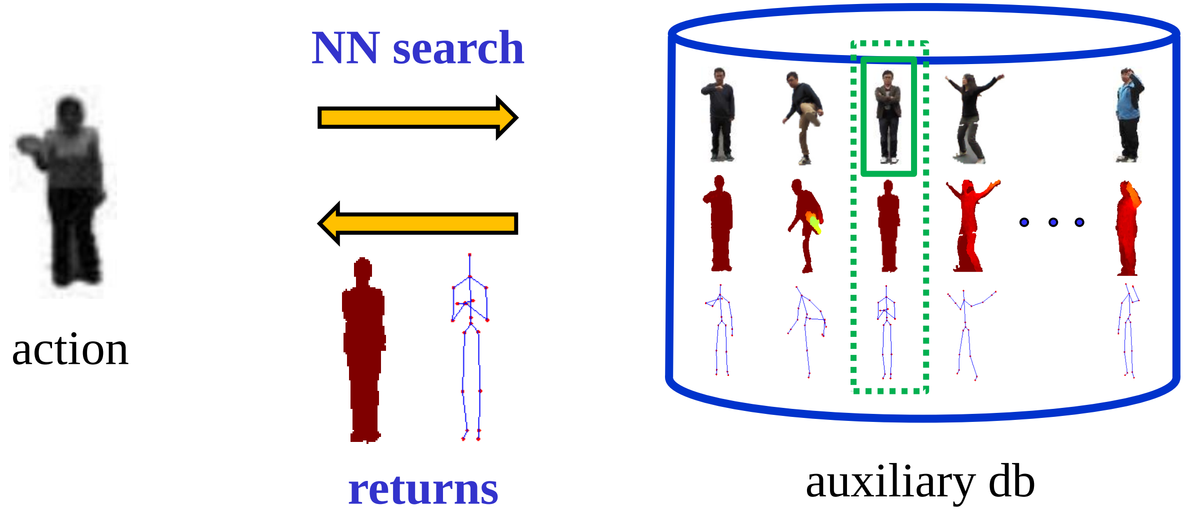
Cross-modal Information Borrowing 1/3

- Fishing 釣漁: cross-modal query expansion



Cross-modal Information Borrowing 2/3

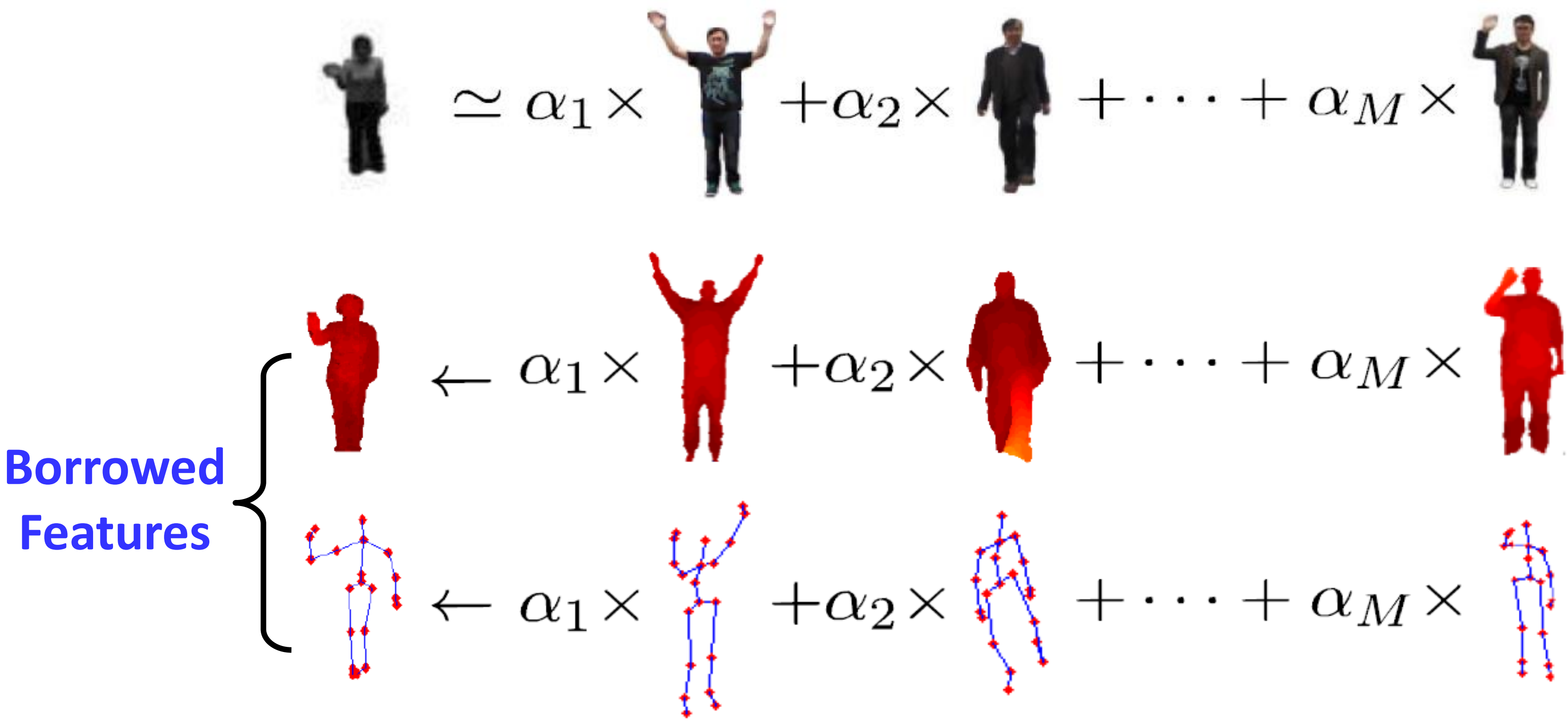
- A naïve way
  - Nearest neighbor search in the RGB domain
  - Borrow the corresponding depth map and skeleton



- It requires a large auxiliary database

Cross-modal Information Borrowing 3/3

- The ``Reconstruct & Borrow'' model



Issues of the reconstruct-&-borrow model

- Domain adaptation
  - Model the variations between the two RGB domains by a linear transformation
- Class-consistent reconstruction coefficients
  - Actions of the same class: similar coefficients
  - Actions of different classes: dissimilar coefficients
- Noisy data or outliers handling
  - Use $\ell_{2,1}$ norm for residual minimization
- Formulate all the three issues into an optimization problem, and solve it

Our approach

- Target database: $D = \{(\mathbf{x}_i, y_i)\}_{i=1}^N$
- Auxiliary database: $\tilde{D} = \{(\tilde{\mathbf{x}}_i, \tilde{\mathbf{d}}_i, \tilde{\mathbf{s}}_i)\}_{i=1}^M$
- Target database augmentation:

$$D = \{(\mathbf{x}_i, y_i)\}_{i=1}^N \Longrightarrow \tilde{D} = \{(\mathbf{x}_i, \mathbf{d}_i, \mathbf{s}_i, y_i)\}_{i=1}^N$$

- Three stages in our approach
  - Domain adaptation
  - Feature augmentation
  - Feature fusion

Domain adaptation

- A reconstruction-based domain adaptation model [*Jhuo et al. CVPR'12*]

transformation

$$W \in \mathbb{R}^{d \times d}$$

recon. coef.

$$[\mathbf{a}_1, \ldots, \mathbf{a}_N] \in R^{M \times N}$$

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

target actions

$$X = [\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N]$$

auxiliary actions

$$\tilde{X} = [\tilde{\mathbf{x}}_1, \tilde{\mathbf{x}}_2, \ldots, \tilde{\mathbf{x}}_M]$$

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

Domain adaptation

- A low-rank reconstruction problem

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

  - $\|E\|_{2,1}$: residual minimization and outlier handling
  - $\operatorname{rank}(A)$: regularization
  - $WW^\top = I$: orthonormal constraint

Domain adaptation

- In our case, the labels of training data are available
  - Class-wise rank minimization

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

- Convex relaxation

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

Domain adaptation

- The optimization problem can be solved by Augmented Lagrange Multiplier (ALM) method

---

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

---

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

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

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

Feature augmentation

- For each target action $\mathbf{x}$ in either training or testing set, we seek its reconstruction coefficients by

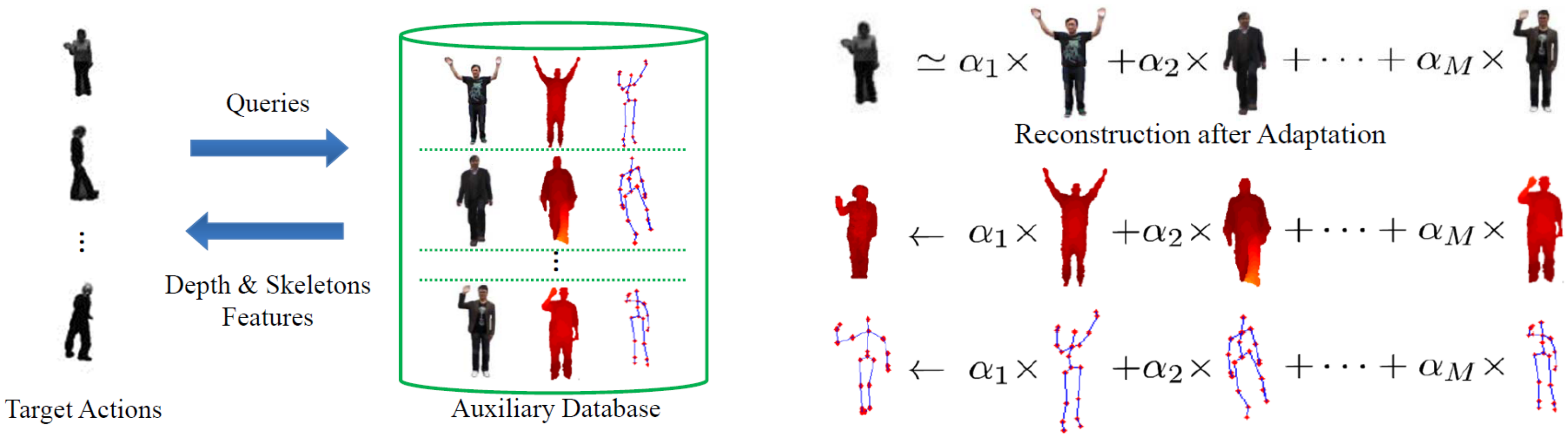
$$\boldsymbol{\alpha} = \arg\min_{\boldsymbol{\alpha}} \|W\mathbf{x} - \tilde{X}\boldsymbol{\alpha}\|^2 + \gamma\|\boldsymbol{\alpha}\|^2$$

- Closed-form solution

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

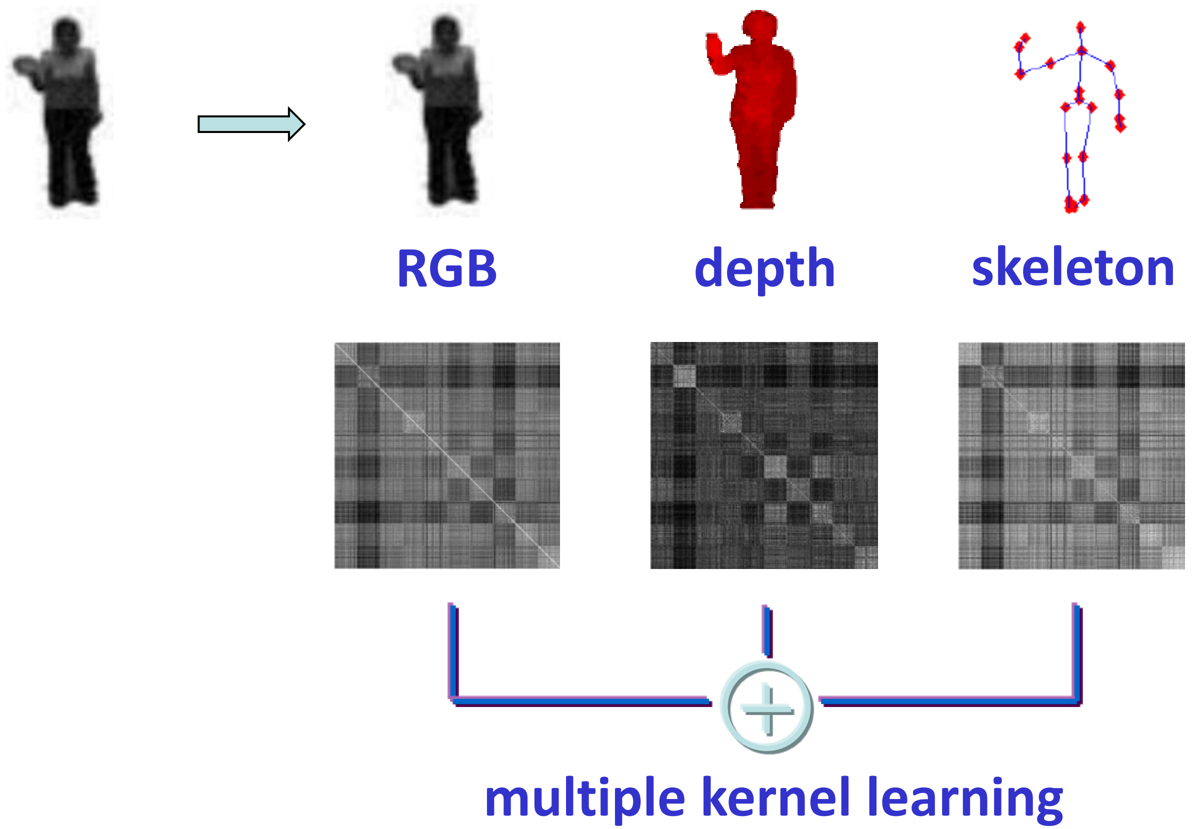
- Feature augmentation $\mathbf{x} \mapsto (\mathbf{x}, \mathbf{d}, \mathbf{s})$ by coefficient sharing
  - Augmented depth map: $\mathbf{d} \leftarrow [\tilde{\mathbf{d}}_1 \cdots \tilde{\mathbf{d}}_M]\boldsymbol{\alpha}$
  - Augmented skeleton: $\mathbf{s} \leftarrow [\tilde{\mathbf{s}}_1 \cdots \tilde{\mathbf{s}}_M]\boldsymbol{\alpha}$
  - For $\mathbf{x}$, how its depth map and skeleton is augmented is the same as how it RGB features are reconstructed

Feature augmentation 2/2



Feature fusion by multiple kernel learning

- Each action is augmented with two borrowed features



Experiments

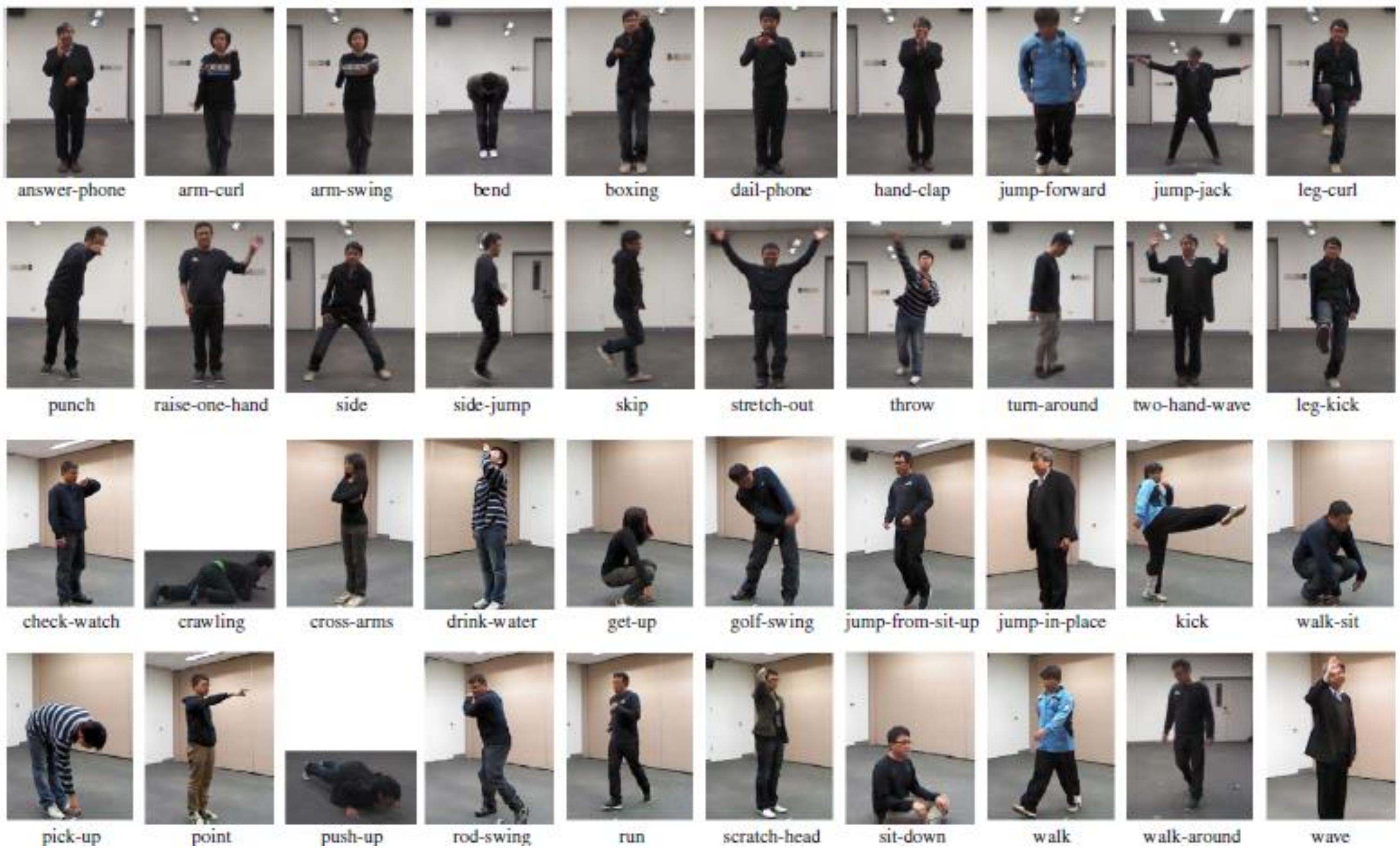
- Three benchmarks of action recognition

	IXMAS	i3DPost	UIUC-1
# classes	11	8	14
# angles of view	3	2	1

- A common auxiliary database
  - Captured by Microsoft Kinect
  - RGB videos
  - Depth maps
  - Skeleton structures

Auxiliary database

- 10 actors, 40 types of actions, 2 views



Video preprocessing and feature representations

- RGB video preprocessing
  - Background estimation [*Tang et al. TMM'12*]
  - Background subtraction [*Barnich et al. TIP'11*]
- RGB videos
  - 3D HOG [*Weinland et al. ICCV'07*]
- Depth maps
  - Spatial-temporal local binary patterns [*Zhao et al. TPAMI'07*]
- Skeleton structures
  - The Fourier temporal pyramid [*Wang et al. CVPR'12*]

Baselines

- **RGB**
  - An SVM classifier that works on $D = \{(\mathbf{x}_i, y_i)\}_{i=1}^{N}$
- **KSDA** (kernel semi-supervised discriminant analysis)
  - Supervised learning on $D = \{(\mathbf{x}_i, y_i)\}_{i=1}^{N}$
  - Manifold regularization on $\tilde{D} = \{\tilde{\mathbf{x}}_i\}_{i=1}^{M}$
- **1NN-Bor**
  - The naïve way for fishing
- **Bor-DEP** & **Bor-SKE**
  - An SVM classifier that works on $D = \{(\mathbf{d}_i, y_i)\}_{i=1}^{N}$
- **Ours**
  - MKL on augmented dataset $D = \{(\mathbf{x}_i, \mathbf{d}_i, \mathbf{s}_i, y_i)\}_{i=1}^{N}$

Experimental results

- LOAO (leave-one-actor-out) cross validation

Method	Ours: d+s	Ours: d	Ours: s	RGB	Bor-DEP	Bor-SKE	KSDA	1NN-Bor	[31]
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.

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

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

Experimental results

- LOAO (leave-one-actor-out) cross validation

Method	Ours: d+s	Ours: d	Ours: s	RGB	Bor-DEP	Bor-SKE	KSDA	1NN-Bor	[31]
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.

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

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

- RGB vs. the state-of-the-art systems

[31] Wu et al. CVPR'11 [12] Iosifidis et al. TNNLS'12 [11] Hernandez et al. Exp. Sys.'13

Experimental results

- LOAO (leave-one-actor-out) cross validation

Method	Ours: d+s	Ours: d	Ours: s	RGB	Bor-DEP	Bor-SKE	KSDA	1NN-Bor	[31]
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.

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

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

- RGB vs. KSDA
- RGB vs. 1NN-Bor

Experimental results

- LOAO (leave-one-actor-out) cross validation

Method	Ours: d+s	Ours: d	Ours: s	RGB	Bor-DEP	Bor-SKE	KSDA	1NN-Bor	[31]
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.

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

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

- RGB vs. Bor-DEP
- RGB vs. Bor-SKE

Experimental results

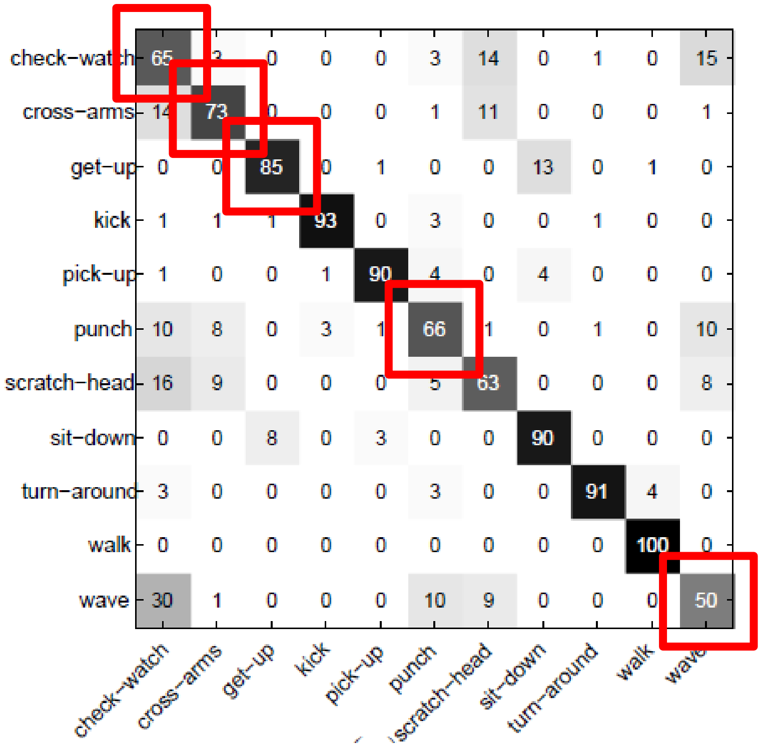
- RGB vs. Ours

	IXMAS	i3DPost	UIUC-1
RGB	78.6%	82.0%	92.1%
Ours (RGB + DEP + SKE)	**89.1%**	**88.3%**	**99.4%**

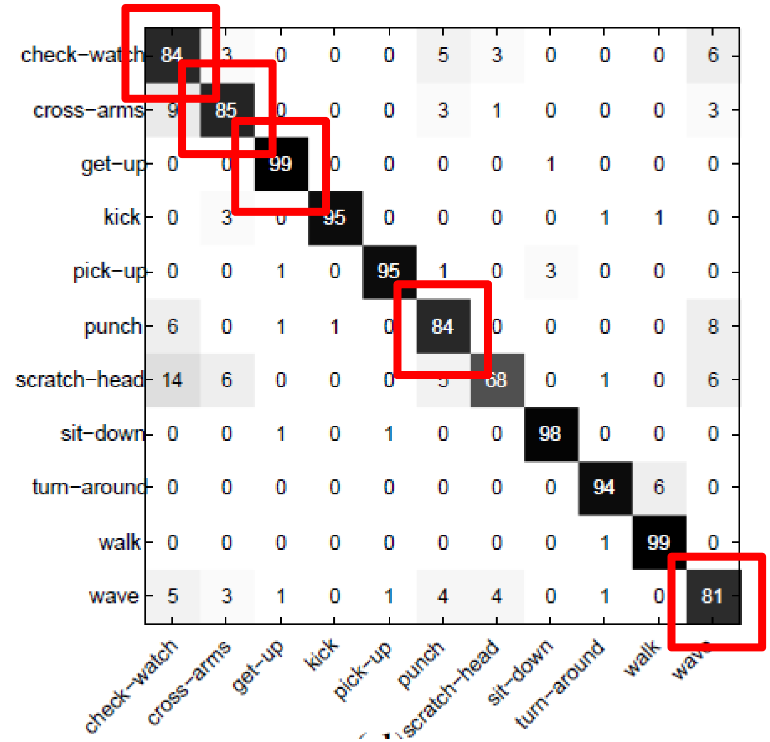
- Performance gains are between 7% ~ 10%
  - Appropriate depth and skeleton features are retrieved
  - MKL determines the effective combinations of features

Experimental results

- Confusion table on IXMAS dataset



RGB



Ours

Conclusions

- Develop new CV techniques with emerging cameras
- A new problem and its solution for addressing the short effective distances of RGB-D cameras
- Fishing: borrowing information from an offline collected, multi-modal database
  - Perform domain adaptation, feature augmentation and fusion
  - Lead to remarkable performance boost on three benchmarks
  - It can be applied to other applications, such as gesture recognition and scene understanding

Thank You for Your Attention!

Yen-Yu Lin (林彥宇)
Email: yylin@citi.sinica.edu.tw