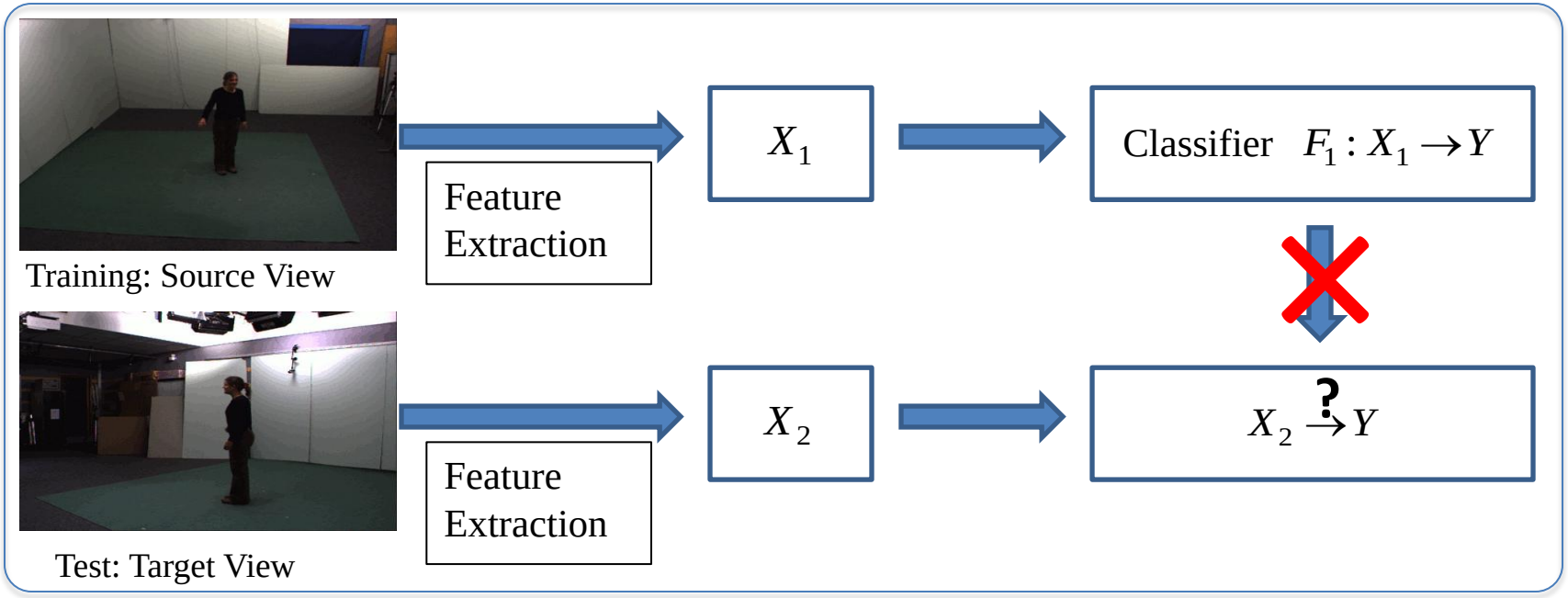
Cross-View Action Recognition via a Transferable Dictionary Pair

Jingjing Zheng[1]
Zhuolin Jiang[1]
Jonathon Phillips[2]
Rama Chellappa[1]

[1]Center for Automation Research, UMIACS, University of University of Maryland, College Park, USA
[2]National Institute of Standards and Technology, USA

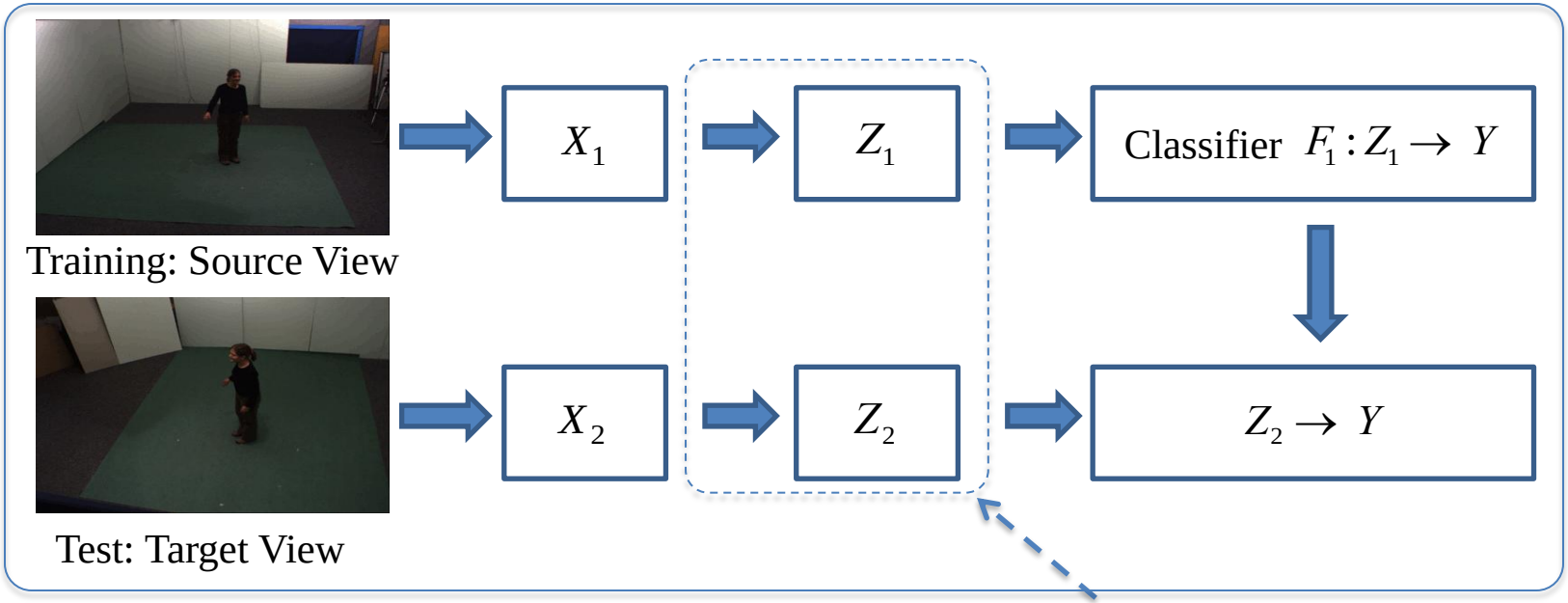
1

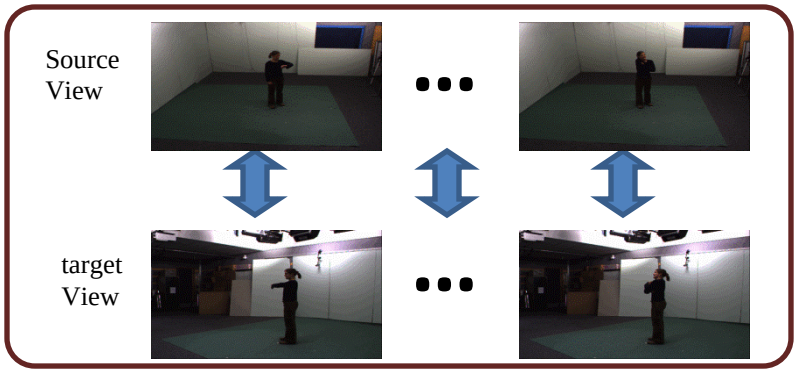
Cross-view Action Recognition



- Directly use classifier F1 trained from view 1 to recognize unknown actions of view 2
  - Performance decrease drastically
  - Motion appearance looks very differently across views

Our Framework





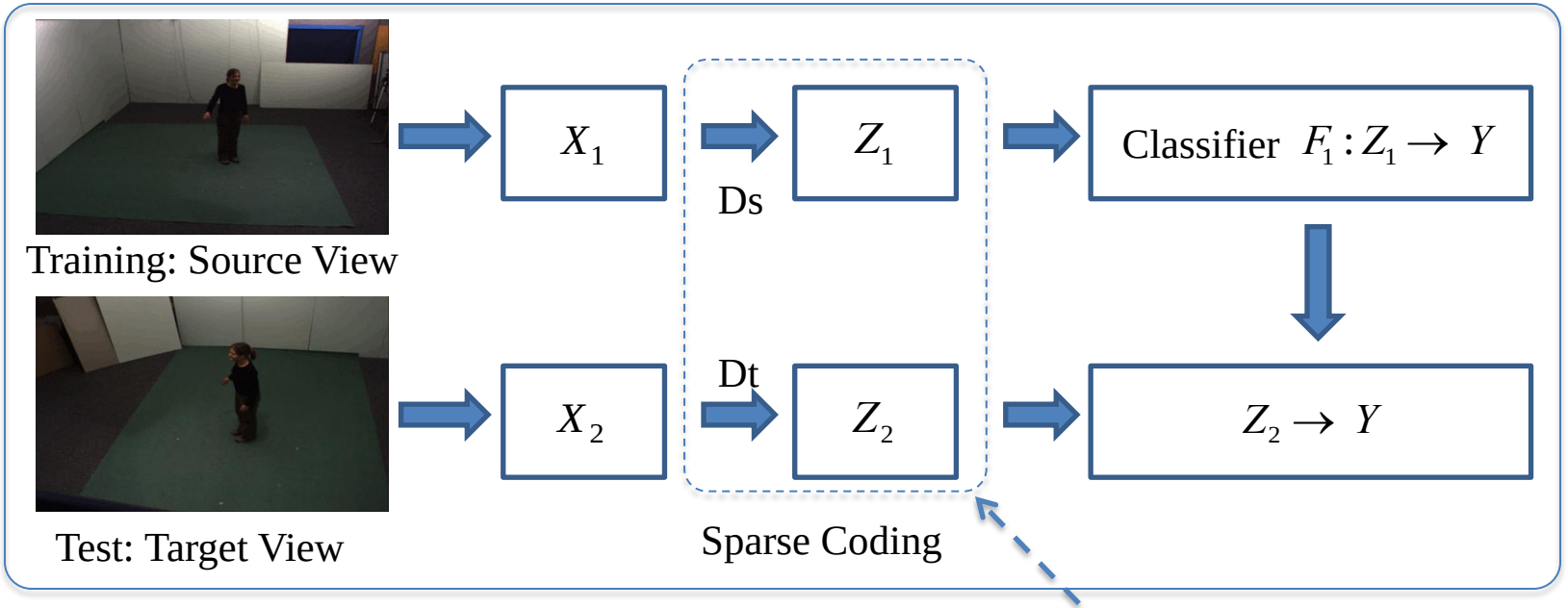
?

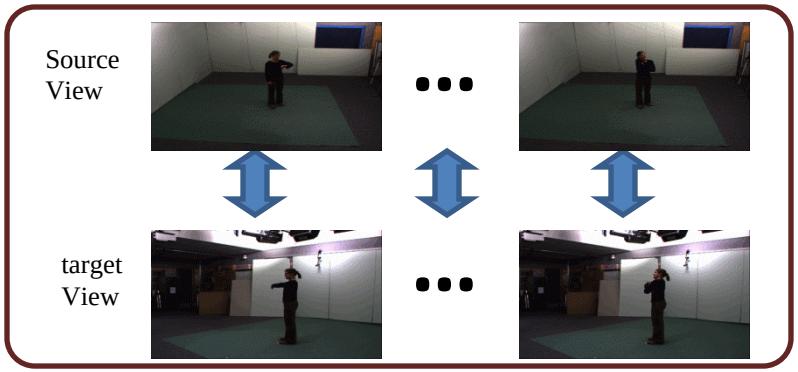
View-invariant Feature Representation

Shared Actions

Shared actions are exclusive from test actions

Our Framework

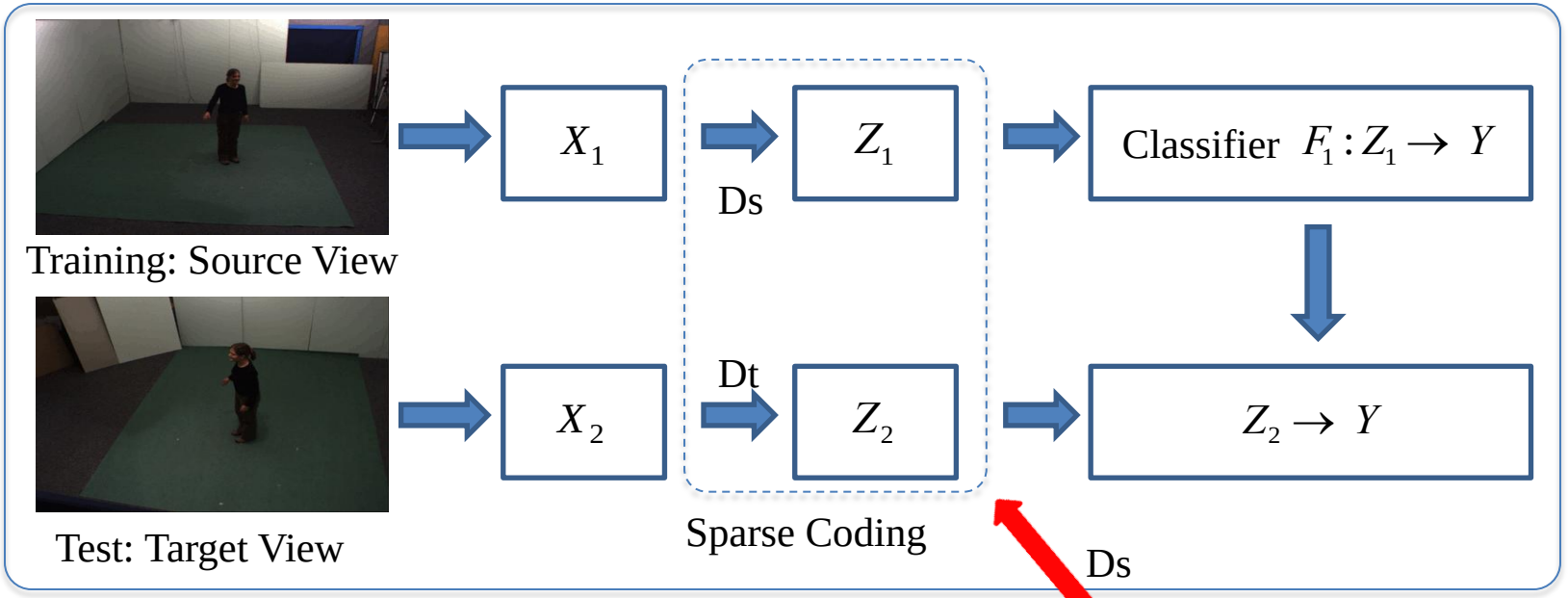


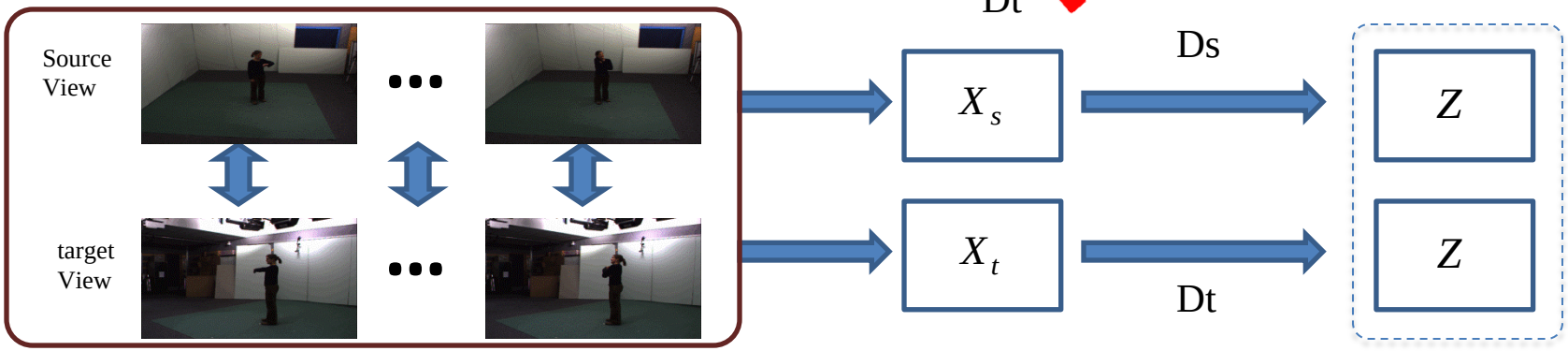


Dictionary Pair {Ds, Dt}

Shared Actions *Encourage each video in a pair to have the same sparse codes*

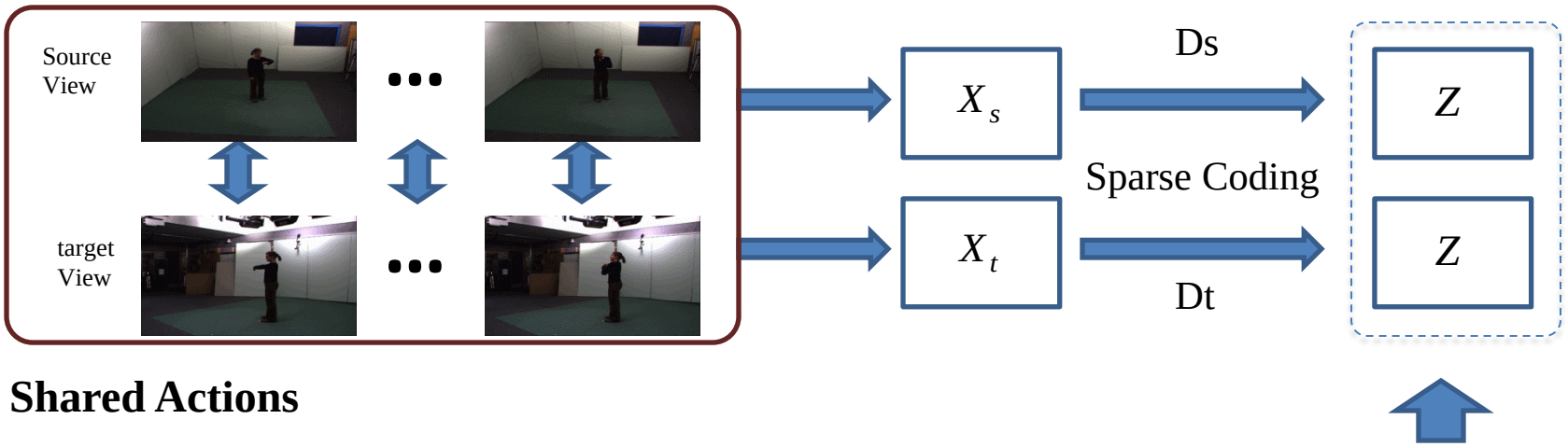
Our Framework





Shared Actions *Encourage each video in a pair to have the same sparse codes*

Our Framework



- *Goal: Encourage two videos in a pair to have the same sparse representations when encoded their corresponding view-dictionary*

- Two settings for learning a transferable dictionary Pair
  - Unsupervised setting ---videos of shared actions are unlabeled
  - Supervised setting --- videos of shared actions are labeled

Review of Dictionary Learning

- K-SVD

Let $X = [x_1, \ldots, x_N] \in \mathrm{R}^{n \times N}$ be a set of n-dim input signals, dictionary $D \in \mathrm{R}^{n \times K}$ $(K > n)$ and sparse codes $Z = [z_1, \ldots, z_N] \in \mathrm{R}^{K \times N}$ is learned by

$$(D, Z) = \arg\min_{D,Z} \|X - DZ\|_2^2 \quad \text{s.t. } \forall i, \|z_i\|_0 \leq s$$

reconstruction error — sparsity constraint

- OMP: Sparse Coding

Given D, the sparse representation $Z = [z_1, \ldots, z_N] \in \mathrm{R}^{K \times N}$ is

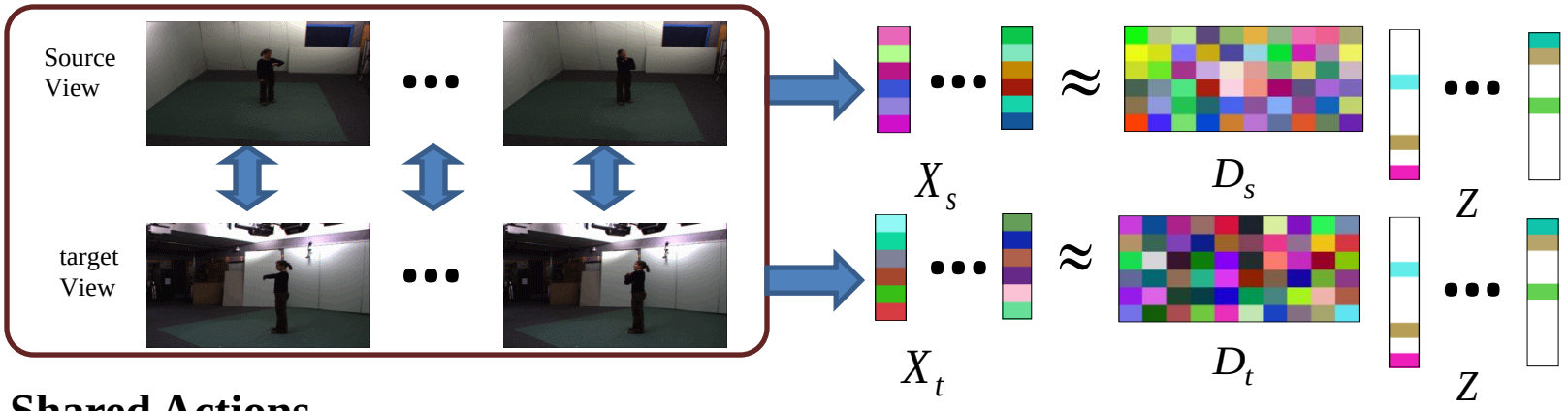
$$Z = \arg\min_{Z} \|X - DZ\|_2^2 \quad \text{s.t. } \forall i, \|z_i\|_0 \leq s$$

Unsupervised Transferable Dictionary Pair Learning

- *Goal: Find discriminative representations that are the same for different views of the same action*

- The objective function of the unsupervised setting:

$$\arg\min_{D_s, D_t, Z} \underbrace{\|X_s - D_s Z\|_2^2 + \|X_t - D_t Z\|_2^2}_{\text{Reconstruction error}}, \quad \text{s.t. } \forall i, \|z_i\| \le s$$



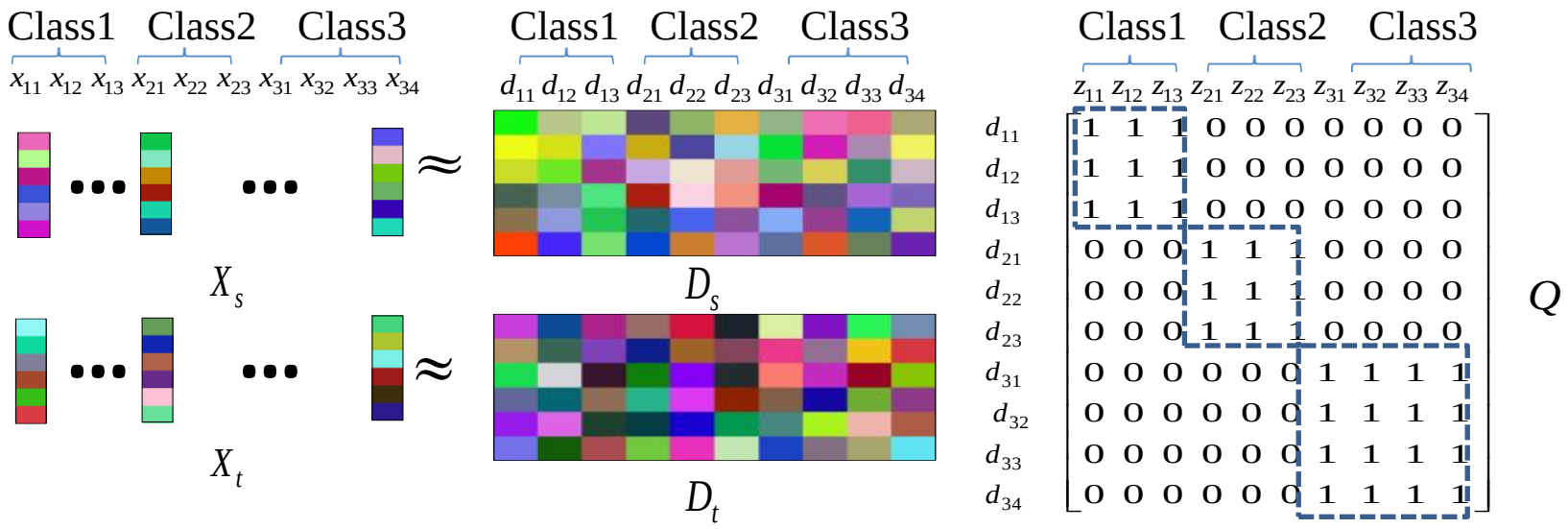
Shared Actions

Supervised dictionary pair learning

- The objective function of the supervised setting:

$$\arg\min_{D_s,D_t,Z} \underbrace{\|X_s - D_s Z\|_2^2 + \|X_t - D_t Z\|_2^2}_{\text{Reconstruction error}} + \lambda \underbrace{\|Q - AZ\|}_{\text{Discriminative sparse code error [1]}}, \quad \text{s.t. } \forall i, \|z_i\| \le s$$

Where A is a linear transformation matrix and Q are "ideal" discriminative sparse codes for the pairs of videos



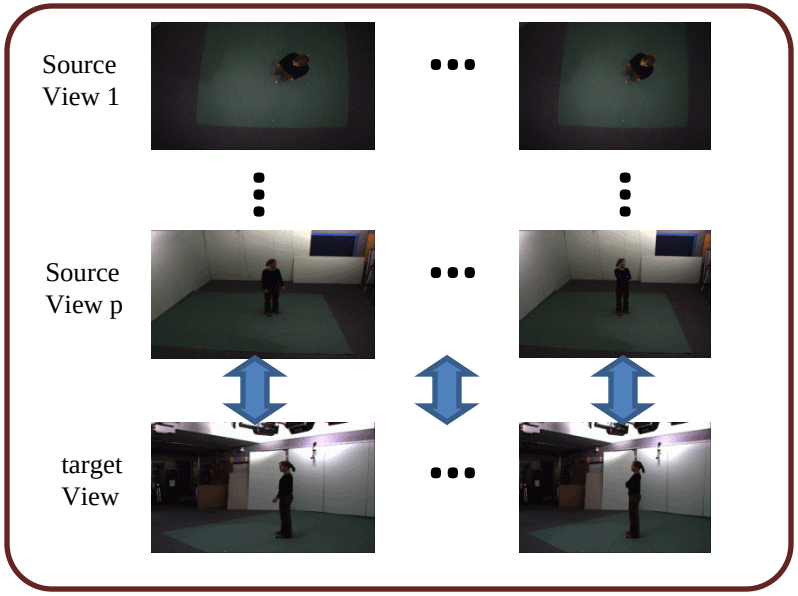
[1] Zhuolin Jiang, Zhe Lin, Larry S. Davis. " Learning a Discriminative Dictionary for Sparse Coding via Label Consistent K-SVD". IEEE Conference on Computer Vision and Pattern Recognition, 2011

Extension: multi-view action recognition

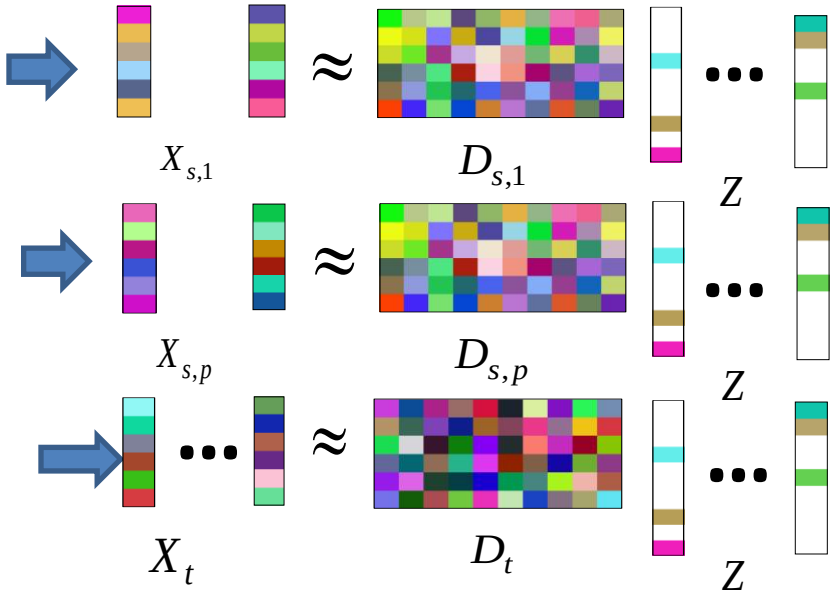
- Assuming p source views and one target view, the objective function is given by

$$\arg \min_{\{D_{s,i}\}_{i=1}^{p}, D_t, Z} \sum_{i=1}^{p} \left\| X_{s,i} - D_{s,i} Z \right\|_2^2 + \left\| X_t - D_t Z \right\|_2^2 \quad \text{s.t.} \quad \forall i, \left\| z_i \right\|_0 \le s$$

Reconstruction error



Shared Actions (Unlabeled)



Experiment: IXMAS multi-view dataset

- Exemplar frames from IXMAS multi-view dataset

- Local feature: STIP feature + Bag of Words (dimension is 1000)
- Global feature: Shape-flow descriptors + Bag of Words (dimension is 500)
- Evaluation strategy: leave-one-action-class-out strategy for choosing the test action
- Classifier: nearest neighbor classifier + L2 norm

Experiment Results

- k-NN without transfer : Independent dictionary pair learning + k-NN
- k-NN with transfer: Transferable dictionary pair learning + k-NN

	Camera0			Camera1			Camera2			Camera3			Camera4		
Cam0				26.4	96.7	98.8	24.6	97.9	99.1	20.3	97.6	99.4	27.9	84.9	92.7
Cam1	31.2	97.3	98.8				23.0	96.4	99.7	23.0	89.7	92.7	20.3	81.2	90.6
Cam2	23.3	92.1	99.4	20.9	89.7	96.4				13.0	94.7	97.3	17.9	89.1	95.5
Cam3	9.7	97.0	98.2	24.9	94.2	97.6	23.0	96.7	99.7				16.7	83.9	90.9
Cam4	51.2	83.0	85.8	38.2	70.6	81.5	41.2	89.7	93.3	53.3	83.7	83.9			
Avg.	28.9	92.4	95.5	27.6	87.8	93.6	28.0	95.1	98.0	27.4	91.2	93.3	20.7	84.8	92.4

Table 1. The accuracy numbers in the bracket are the average recognition accuracies of k-NN without transfer (in black), our unsupervised and supervised approaches (in red).

Experiment Results

- Cross-view action recognition of unsupervised dictionary pair learning

	Camera0	Camera1	Camera2	Camera3	Camera4
Cam0		72 77.6 79.9 96.7	61 69.4 76.8 97.9	62 70.3 76.8 97.6	30 44.8 74.8 84.9
Cam1	69 77.3 81.2 97.3		64 73.9 75.8 96.4	68 67.3 78.0 89.7	41 43.9 70.4 81.2
Cam2	62 66.1 79.6 92.1	67 70.6 76.6 89.7		67 63.6 79.8 94.9	43 53.6 72.8 89.1
Cam3	63 69.4 73.0 97.0	72 70.0 74.1 94.2	51 51.8 74.0 96.7		44 44.2 66.9 83.9
Cam4	51 39.1 82.0 83.0	55 38.8 68.3 70.6	61 51.8 74.0 89.7	53 34.2 71.1 83.7	
Avg.	61 63.0 79.0 92.4	67 43.3 74.7 87.8	61 64.5 75.2 95.1	63 58.9 76.4 91.2	40 46.6 71.2 84.8

Table 1. The accuracy numbers in the bracket are the average recognition accuracies of three state-of-art approaches(in black) and our unsupervised approaches (in green).

Experiment Results

- Cross-view action recognition of supervised dictionary pair learning

	Camera0	Camera1	Camera2	Camera3	Camera4
Camera0		79 98.8	79 99.1	68 99.4	76 92.7
Camera1	72 98.8		74 99.7	70 92.7	66 90.6
Camera2	71 99.4	82 96.4		76 97.3	72 95.5
Camera3	75 98.2	75 97.6	73 99.7		76 90.0
Camera4	80 85.5	73 81.5	73 93.3	79 83.9	
Avg.	74 95.5	77 93.6	76 98.0	73 93.3	72 92.4

Table 2. The accuracy numbers in the bracket are the average recognition accuracies of one state-of-the-art approach (Farhadi et al. ICCV 2009 in black) and our supervised approaches (in green).

Experiment Results

- Multi-view action recognition

	Camera0	Camera1	Camera2	Camera3	Camera4	Avg.
Our unsupervised	98.5	99.1	99.1	100	90.3	97.4
Our supervised	99.4	98.8	99.4	99.7	93.6	98.2
LWE	86.6	81.1	80.1	83.6	82.8	82.8
Junejo et. al.	74.8	74.5	74.8	70.6	61.2	71.2
Liu et. al.	76.7	73.3	72.0	73.0	N/A	73.8
Weinland et. Al.	86.7	89.9	86.4	87.6	66.4	83.4

Table 3. Multi-view action recognition results. Each column corresponds to one target view.

Experiment Results

- Multi-view action recognition

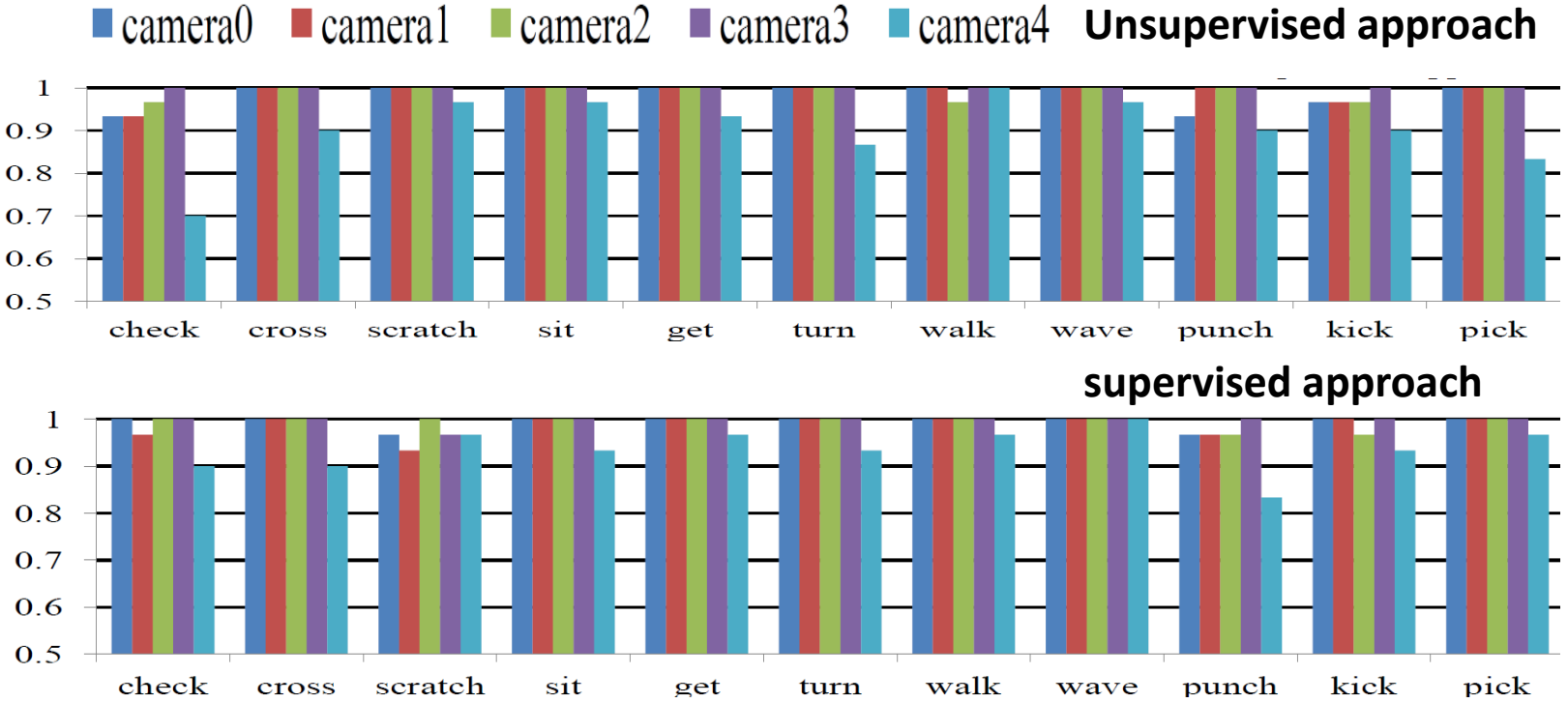


Figure 1. The multi-view recognition results on each action category.

Note: It is harder to transfer action models from across views that involves the top view.

Conclusions

- Directly exploits the video-level correspondence
- Bridge the gap of sparse representations of pairs of videos taken from different views of the same action.
- Can be applied to multi-view action recognition
- Achieves state-of-the-art performance

Thank You!

- Acknowledgements:

- University of Maryland, Computer Vision Lab
- NIST