Enhancing Action Recognition by Cross-Domain Dictionary Learning

Fan Zhu, Ling Shao

Regular action recognition:

sufficient training samples are available.

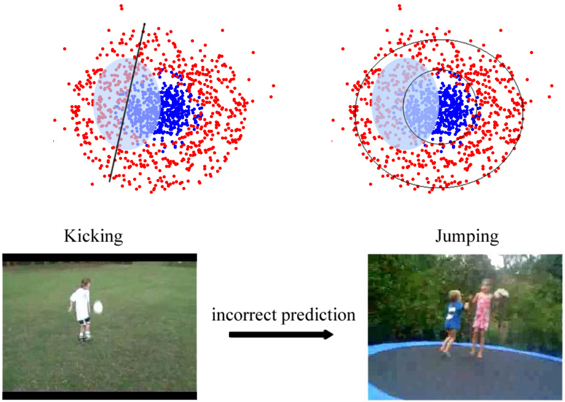
The scenario we are trying to address:

only a few training samples that stay in the same feature space or share the same distribution with the testing data are available

Reasons:

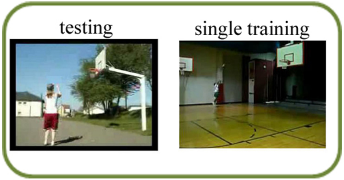
- high price of human manual annotation
- environmental restrictions

Problem with insufficient training data:



Typical examples:

- one-shot sequence learning (*Fei-Fei et al. CVPR'07*).

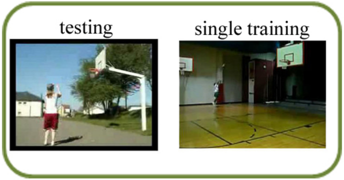


- cross-view action recognition (*Zheng et al. BMVC'13*).

What shall we do when no sufficient training data are available?

Typical examples:

- one-shot sequence learning (*Fei-Fei et al. CVPR'07*).
- cross-view action recognition (*Zheng et al. BMVC'13*).



What shall we do when no sufficient training data are available?

- expensive training data
- expensive algorithms

We show two facts of the human vision system in advanced to answering this question:

- The first fact: humans are able to learn tens of thousands of visual categories in their life.
- The second fact: humans' visual impressions towards the same action or the same object cover a wide range. (e.g., an action seen from 2D static images vs. the same action seen from 3D dynamic movies or an object seen from real-world scenes vs. the same object seen from low-resolution online images.)

Based on these two facts, we introduce a new action recognition framework, which

- utilizes relevant actions from other domains as auxiliary knowledge (motivated by the first fact).
- spans the intra-class diversity of the original learning system (motivated by the second fact).

Typical setting in transfer learning:

In transfer learning, both the training data and the testing data can contribute to two types of domains:

- the target domain: contains the testing instances,and a few training instances.
- the source domain: contains training instances.

Typical transfer learning algorithms:

- Adaptive Support Vector Machines (*Yang et al. ICM'07*).
- Transfer Multiple Kernel Learning (*Duan et al. PAMI'12*).
- TrAdaBoost (*Dai et al. ICML'07*).

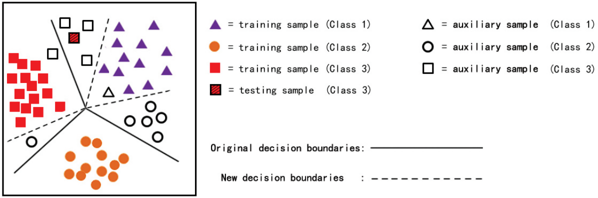


Figure 1: Illustration of how the categorization system can gain more discriminative power through the collaboration with the source domain data in the 2-dimensional feature space.

.

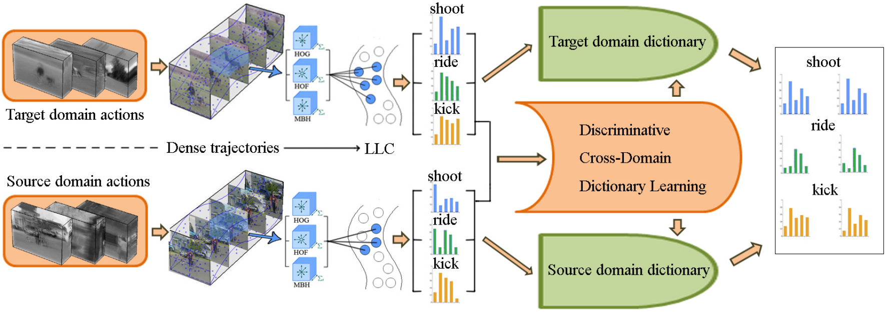


Figure 2: Outline of the proposed framework.

.

Dictionary learning:

Let D_t be the target domain dictionary, Y_t and X_t be the set of target domain input signals and the set of sparse signals. Learning a reconstructive dictionary for obtaining the sparse representation of the target domain signals can be accomplished by solving the following optimization problem:

$$< D_t, X_t >= \arg \max_{D_t, X_t} \underbrace{\|Y_t - D_t X_t\|_2^2}_{reconstruction\ error} \quad \underbrace{s.t. \forall i, \|x_t^i\|_0 \leq T}_{sparsity\ constraint}, \tag{1}$$

where T is the sparsity constraint factor. Similarly, the source domain dictionary and sparse signals can be obtained through:

$$< D_s, X_s >= \arg \max_{D_s, X_s} \|Y_s - D_s X_s\|_2^2 \quad s.t. \forall i, \|x_s^i\|_0 \leq T, \tag{2}$$

The goal:

to force the mismatched sparse representations from different domains into the same feature space, so that the smoothness property can be satisfied in the new feature space.

$$\begin{aligned} \langle D_t, D_s, X_t, X_s \rangle = \arg \min_{D_t, D_s, X_t, X_s} \|Y_t - D_t X_t\|_2^2 \\ + \|Y_s - D_s X_s\|_2^2 + \|X_t - f(Y_t, Y_s) X_s\|_F^2 \\ + \|X_s - f(Y_s, Y_t) X_t\|_F^2 \\ s.t. \forall i, [\ \|x_t^i\|_0, \|x_s^i\|_0\ ] \leq T, \end{aligned} \tag{3}$$

where the function $f(\cdot)$ computes the mapping of correspondence samples across different domains.

$$\mathbb{A}_1 = \begin{pmatrix} \Psi(y_t^1, y_s^1) & \cdots & \cdots & \Psi(y_t^1, y_s^{c_s^1}) \\ \vdots & \ddots & & \vdots \\ \vdots & & \ddots & \vdots \\ \Psi(y_t^{c_t^1}, y_s^1) & \cdots & \cdots & \Psi(y_t^{c_t^1}, y_s^{c_s^1}) \end{pmatrix}, \quad (4)$$

where $\Psi(y_t^i, y_s^j)$ in each $\mathbb{A}_c$ can be computed by the Gaussian kernel.

$$\mathbb{A}_c(i,j) = \begin{cases} 1, & if \quad \mathbb{A}_c(i,j) = max(\mathbb{A}_c(:,j)) \\ 0, & otherwise. \end{cases} \quad (5)$$

and similarly for category 2:

$$\mathbb{A}_2 = \begin{pmatrix} \Psi(y_t^{c_t^1+1}, y_s^{c_s^1+1}) & \cdots & \cdots & \Psi(y_t^{c_t^1+1}, y_s^{c_s^2}) \\ \vdots & \ddots & & \vdots \\ \vdots & & \ddots & \vdots \\ \Psi(y_t^{c_t^2}, y_s^{c_s^1+1}) & \cdots & \cdots & \Psi(y_t^{c_t^2}, y_s^{c_s^2}) \end{pmatrix}, \tag{6}$$

$$\mathbb{A} = \begin{pmatrix} \mathbb{A}_1 & & & \\ & \mathbb{A}_2 & & \\ & & \ddots & \\ & & & \mathbb{A}_C \end{pmatrix}, \tag{7}$$

$$\begin{aligned}\langle D_t, D_s, X_t, X_s\rangle = \arg \min_{D_t, D_s, X_t, X_s} &\|Y_t - D_t X_t\|_2^2 + \|Y_s - D_s X_s\|_2^2 \\ &+ \underbrace{\|X_t - f(Y_t, Y_s)X_s\|_F^2 + \|X_s - f(Y_s, Y_t)X_t\|_F^2}_{disappears\ under\ the\ perfect\ mapping\ assumption} \\ &s.t. \forall i, [\ \|x_t^i\|_0, \|x_s^i\|_0\ ] \leq T,\end{aligned}$$

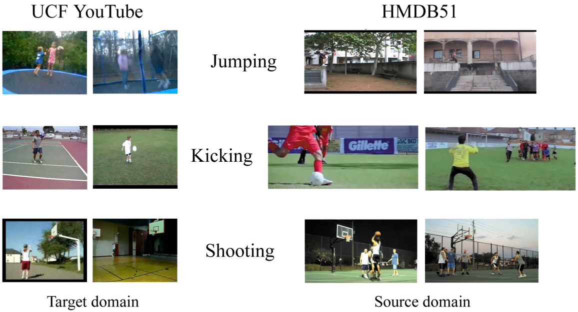
$$\begin{aligned}\langle D_t, D_s, X_t, X_s\rangle = arg \min_{D_t, D_s, X_t} \|Y_t - D_t X_t\|_2^2 \\ + \|(\mathbb{A} Y_s^T)^T - D_s X_t\|_2^2 \quad s.t. \forall i,\ \|x_t^i\|_0 \leq T.\end{aligned} \tag{8}$$

We attempt to further include a discriminative term to the objective function with respect to the optimal data distribution. Let the classifier $\mathcal{F}(x)$ satisfy the following equation:

$$\mathcal{P} = arg\min_{\mathcal{P}} \sum_i w_i \times \mathcal{L}\{h_i, \mathcal{F}(x_t^i, \mathcal{P})\} + \lambda_i \|\mathcal{P}\|_F^2, \tag{9}$$

$$\begin{aligned}\langle D_t, D_s, X_t, \Phi, \mathcal{P}\rangle = \underbrace{arg \min_{D_t, D_s, X_t, \Phi, \mathcal{P}} \|Y_t - D_t X_t\|_2^2 + \|Y_s \mathbb{A}^T - D_s X_t\|_2^2}_{\text{reconstruction error}} \\ + \underbrace{\alpha\|Q - \Phi X_t\|_2^2 + \beta\|\mathcal{H} - \mathcal{P} X_t\|_2^2}_{\text{discriminative sparse code error}} \quad s.t. \forall i,\ \|x_t^i\|_0 \leq T,\end{aligned} \tag{10}$$

Experiments: UCF YouTube dataset + HMDB51 dataset



Representation: dense trajectories + LLC coding.

Vocabulary size: 4000.

Table 1: Performance comparison between DCDDL and other methods on the UCF YouTube dataset.

Algorithm	LLC	LLC	K-SVD	K-SVD	LC-KSVD	LC-KSVD	DCDDL
Learning	N/A	N/A	Un	Un	Su	Su	Su
Source data	No	Yes	No	Yes	No	Yes	Yes
24 actors	86.67%	86.67%	82.22%	77.78%	86.67%	82.22%	**88.89**%
20 actors	75.42%	70.21%	68.75%	72.08%	75.42%	75.42%	**77.50**%
16 actors	70.88%	70.17%	63.96%	67.54%	72.08%	72.08%	**73.03**%
09 actors	61.41%	61.80%	55.70%	59.15%	65.25%	64.72%	**66.31**%
05 actors	54.10%	53.35%	50.05%	48.88%	56.55%	54.10%	**56.66**%

Table 2: Performance comparison under the leave-one-actor-out setting.

Methods	[1]	[2]	BoF [3]	DCDDL
Results	71.2%	75.21%	80.02%	**82.52%**

References

[1] J. Liu, J. Luo, and M. Shah. Recognizing realistic actions from videos "in the wild". In *CVP*. 2009.

[2] N. Ikizler-Cinbis and S. Sclaroff. Object, scene and actions: Combining multiple features for human action recognition. In *ECCV*. 2010.

[3] H. Wang, A. Klaser, C. Schmid, and C. L. Liu. Action recognition by dense trajectories. In *CVP*. 2011.

Conclusion:

- We present a cross-domain action recognition framework that attempts to enhance the performance of the original recognition system by spanning the intra-class diversities of the target domain training actions.
- The proposed discriminative cross-domain dictionary learning technique copes with the feature distribution mismatch problem.
- Achieves state-of-the-art performance
- Can be adapted to solve many real-world transfer learning problems.

Thank you!