Enhancing Action Recognition by Cross-Domain Dictionary Learning

Fan Zhu
fan.zhu@sheffield.ac.uk
Ling Shao
ling.shao@sheffield.ac.uk

Department of Electronic and Electrical Engineering
The University of Sheffield
Sheffield, S1 3JD, UK

Our work is inspired by two facts of the human vision system. The first fact is that humans are able to learn tens of thousands of visual categories in their life, which leads to the hypothesis that humans achieve such a capability by accumulated information and knowledge. Another fact is that human's visual impressions towards the same action or the same object comes from 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. These facts can be explained in the computer vision language as the human vision system possesses the ability of spanning the intra-class diversity of the original training instances through transferring prior knowledge. Motivated by the above two facts, we introduce a new action recognition framework that utilizes relevant actions from other domains as auxiliary knowledge (motivated by the first fact) to span the intra-class diversity of the original learning system (motivated by the second fact). In addition to manually annotated actions in the target domain, labeled actions from a different domain are provided as the source domain actions. Based on the recent success of dictionary learning methods in solving computer vision problems, we present a discriminative cross-domain dictionary learning (DCDDL) technique to learn a reconstructive, discriminative and domain-adaptive dictionary pair for data under different distributions. The flowchart of the proposed framework is shown in Figure 1.

The objective function:

$$\begin{aligned}\langle D_t, D_s, X_t, X_s \rangle = arg \min_{D_t,D_s,X_t,X_s} \|Y_t - D_t X_t\|_2^2 \\ + \|Y_s - D_s X_s\|_2^2 + \|X_t - f(Y_t,Y_s)X_s\|_F^2 \\ + \|X_s - f(Y_s,Y_t)X_t\|_F^2 \\ s.t. \forall i, [\ \|x_t^i\|_0, \|x_s^i\|_0\ ] \leq T,\end{aligned} \tag{1}$$

where the function $f(\cdot)$ computes the mapping of correspondence samples (i.e., samples that share the same class labels while being close to each other) across different domains. Thus, small values of $\|X_t - f(Y_t,Y_s)X_s\|_F^2$ and $\|X_s - f(Y_s,Y_t)X_t\|_F^2$ indicate that those data points close to each other are more likely to share the same class label in the new target feature domain and the new source feature domain respectively. Since we are only concerned with the smoothness within the target domain data, the last term in Equation (1) can be removed. According to the stated scenario, no manually annotated correspondences between the target domain data and the source domain data are available in the training phase, thus $f(\cdot)$ is computed using a category-specific searching method. We use the matrix $\mathcal{A}$ to represent the connections between the target domain data and the source domain data in the category-specific manner. In order to establish the correspondences across the target domain data and the source domain data, the maximum element in each column of $\mathbb{A}$ is preserved and set to 1 while the remaining elements are set to 0.

$$\mathbb{A}(i,j) = \begin{cases} 1, & if \quad \mathbb{A}(i,j) = max(\mathbb{A}(:,j)) \\ \\ 0, & otherwise. \end{cases} \tag{2}$$

Assuming $\mathbb{A}$ leads to a perfect mapping across the sparse codes from both domains and the matched pair of samples in different domains possesses an identical representation after encoding, the Equation (1) can be rewritten as:

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

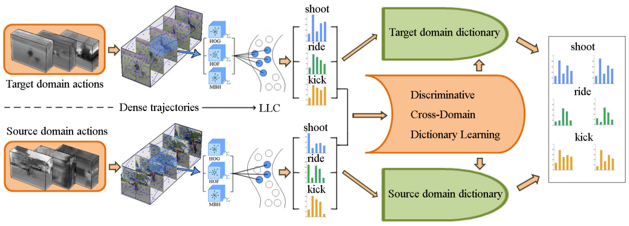

Figure 1: Flowchart of the proposed cross-domain action recognition framework.

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**%

We 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:

$$\mathcal{P} = arg \min_{\mathcal{P}} \sum_i w_i \times \mathcal{L}\{h_i, \mathcal{F}(x_t^i, \mathcal{P})\} + \lambda_i \|\mathcal{P}\|_F^2, \tag{4}$$

where $\mathcal{L}$ is the classification loss function, h_i indicates the target domain labels of x_t^i, $\mathcal{P}$ denotes the classifier parameters and λ_i is a regularization parameter.

To demonstrate the effectiveness of our approach, experiments are conducted using two data sources, where the UCF YouTube action dataset [2] is treated as the target domain and the HMDB51 dataset [1] is treated as the source domain. Specifically, 7 body movements, including ride bike, dive, golf, jump, kick ball, ride horse and shoot ball, are chosen from the HMDB51 dataset in correspondence with similar actions in the UCF YouTube dataset. Performance comparisons of the proposed method and other state-of-the-art methods are reported on both scenarios where the source domain data are included or excluded in Table 1 and Figure 2.

[1] H. Kuehne, H. Jhuang, E. Garrote, T. Poggio, and T. Serre. Hmdb: a large video database for human motion recognition. In *IEEE International Conference on Computer Vision*. 2011.

[2] J. Liu, J. Luo, and M. Shah. Recognizing realistic actions from videos "in the wild". In *IEEE Conference on Computer Vision and Pattern Recognition*. 2009.

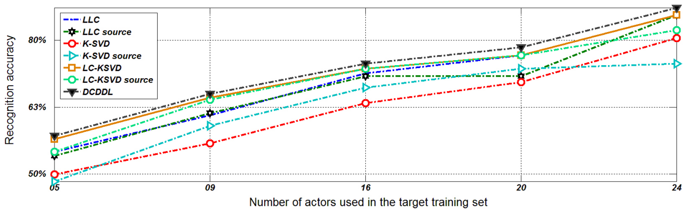

Figure 2: Performance comparison of the proposed DCDDL with other methods under different dataset partitions.