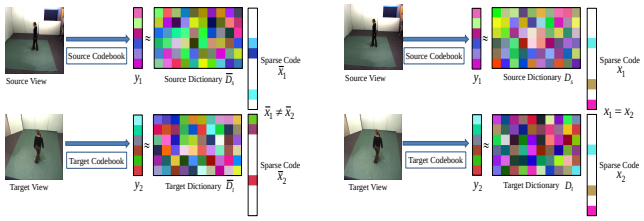
Cross-View Action Recognition via a Transferable Dictionary Pair

Jingjing Zheng[1]
zjngjng@umiacs.umd.edu
Zhuolin Jiang[2]
zhuolin@umiacs.umd.edu
P. Jonathon Phillips[3]
jonathon.phillips@nist.gov
Rama Chellappa[1]
rama@umiacs.umd.edu

[1] Department of Electrical and Computer Engineering and the Center for Automation Research, UMIACS
University of Maryland
College Park, MD, USA

[2] UMIACS, University of Maryland
College Park, MD, USA

[3] National Institute of Standards and Technology
Gaithersburg, MD, USA



(a) Independent dictionary pair (b) Transferable dictionary pair

Figure 1: **Independent dictionary pair versus Transferable dictionary pair.** (a) Based on the BoVW feature representation, the source and target dictionaries are learned individually using videos taken from two different views of the same action. (b) Based on the same BoVW feature representation, we simultaneously learn the source and target dictionaries by forcing the shared videos taken from two views to have the same sparse representations.

In this paper, we propose a novel approach for cross-view action recognition by transferring sparse feature representations of videos from the source to target view. The first step is to construct a separate codebook for each view, where the first view is the source domain and the second is the target domain. Each codebook is constructed by the k-means clustering algorithm. Each video is modeled as a Bag of Visual Words (BOVW) using the corresponding codebook from the same view. Although each pair of videos records the same action from two views, the feature representations of an action in the two views is different because each view has its own codebook. The next step is to learn a dictionary pair $\{D_s, D_t\}$, with D_s corresponding to the source view and D_t the target view. The dictionaries are designed to have sparse codes that are the same for each pair of videos that records the same action across the two views. In this way, videos across different views of the same action are encouraged to have similar sparse representations. This procedure enables the transfer of the sparse feature representations of videos in the source view to the corresponding videos in the target view. There is no reason to assume that two separate dictionaries that are learned independently for each view will have a view-invariant feature representation. The difference between learning a dictionary pair individually and our transferable dictionary pair learning can be seen in Figure 1.

Furthermore, we consider two types of actions: *shared* actions, that are observed in both *source* and *target* views, and *orphan* actions that are observed only in the source view. Orphan action labels are available only in the source view. For the shared actions, we consider two scenarios: (1) shared actions in both views are not labeled; (2) shared actions in both views are labeled. We refer them as the unsupervised and supervised settings respectively and propose corresponding unsupervised and supervised approaches for learning the transferable dictionary pair. Note that under both settings only videos of shared actions across different views are used for learning the dictionary pair, which means that the dictionary pair is not affected by videos of orphan actions.

In the unsupervised setting, our goal is to transfer orphan action models from the source view to the target view. For this purpose, we construct a transferable dictionary pair denoted by $\{D_s, D_t\}$, such that each pair of videos of the same action taken from the source and target views have the same sparse representations. Let $Y_s, Y_t \in \mathbb{R}^{n\times M}$ denote the feature representations of M videos of shared actions in source and target views. The objective function for learning a transferrable dictionary pair is given by:

$$arg \min_{D_s,D_t,X} ||Y_s - D_s X||_2^2 + ||Y_t - D_t X||_2^2 \quad \text{s.t.} \quad \forall i, ||x_i||_0 \leq s. \qquad (1)$$

In supervised setting where action categories of shared action videos are available in both views, we leverage this category information to learn a discriminative transferrable dictionary pair. Here the key idea is to partition the total dictionary items into disjoint subsets and each subset is responsible for representing videos of one action. The intuition behind this idea is that action videos from the same class tend to have same features and each action video could be well represented by other videos from the same class. On the contrary, videos from different classes tend to have different features and thus should be well represented by disjoint subsets of other videos. In order to achieve the above goal, we incorporate a label consistent regularization term introduced in [1] to the objective function in Eq. 1. Now the objective function for dictionary pair construction is given by:

$$arg \min_{D_s,D_t,A,X} ||Y_s - D_s X||_2^2 + ||Y_t - D_t X||_2^2 + \lambda ||Q - AX||_2^2 \quad \text{s.t.} \quad \forall i, ||x_i||_0 \leq s, \qquad (2)$$

where λ controls the tradeoff between the reconstruction error and label consistent regularization. The elements of matrix $Q = [q_1, ..., q_N] \in \mathbb{R}^{K\times N}$ are made of the ideal "discriminative" sparse codes of shared action videos in both views. The vector $q_i = [q_i^1, ..., q_i^K] = [0...1, 1, ..0] \in \mathbb{R}^K$ is a discriminative sparse code corresponding to one shared action video pair $\{y_{s,i}, y_{t,i}\}$ and the non-zeros values of q_i occur at those indices where the shared action video pair $\{y_{s,i}, y_{t,i}\}$ and the dictionary item d_k share the same label. Thus matrix A is a linear transformation matrix which transforms the original sparse code X to be most discriminative in sparse feature space $\mathbb{R}^K$.

In order to handle the situation where videos of shared actions across multiple source views are available, we propose to learn a set of view-dependent dictionaries by forcing videos of shared actions in all views to have the same representations when encoded using the corresponding view-dependent dictionary. Suppose there are p source views $\mathcal{V}^s$ and one target view $\mathcal{V}^t$, the corresponding objective function is given by:

$$arg \min_{\{D_{s,i}\}_{i=1}^p, D_t, X} \sum_{i=1}^{p} ||Y_{s,i} - D_{s,i} X||_2^2 + ||Y_t - D_t X||_2^2 \quad \text{s.t.} \quad \forall i, ||x_i||_0 \leq s. \qquad (3)$$

Given the learned view specific dictionaries, we obtain the sparse representation of each video in each view using the corresponding view-dependent dictionary. Videos of orphan actions in different views will have similar sparse representations when encoded using the corresponding view-dependent dictionary. This is because dictionaries are learned by forcing different sets of videos of shared actions in different views to have the same sparse representations. Thus, the action model learned in one view can be directly applied to classify unlabeled test videos in another different view.

We have extensively tested our approach on the publicly available IXMAS multi-view dataset [2]. The resulting performance clearly confirms the effectiveness of our approach for cross-view action recognition.

[1] Zhuolin Jiang, Zhe Lin, and Larry S. Davis. Learning a discriminative dictionary for sparse coding via label consistent K-SVD. In *CVPR*, 2011.

[2] Daniel Weinland, Edmond Boyer, and Rémi Ronfard. Action recognition from arbitrary views using 3D exemplars. In *ICCV*, 2007.