Learning a Non-linear Knowledge Transfer Model for Cross-View Action Recognition

Hossein Rahmani, Ajmal Mian
Computer Science and Software Engineering, The University of Western Australia.

Action recognition from videos is a significant research problem with applications in human-computer interaction, smart surveillance, and video retrieval. Several techniques have been proposed for discriminative action representation such as 2D shape matching, spatio-temporal interest points, and trajectory-based representation. While these methods are effective for action recognition from a common viewpoint, their performance degrades significantly under viewpoint changes. This is because the same action appears quite different when observed from different viewpoints.

A practical system should be able to recognize human actions from different unknown and more importantly unseen views. Recently, knowledge transfer-based methods [1, 4, 5, 6] have become popular for cross-view action recognition. These methods find a view independent latent space in which features extracted from different views are directly comparable. Such methods only seek a set of linear transformations connecting source and target views (see Fig. 1) and are thus unable to capture the non-linear manifolds where realistic action videos usually lie on, especially when actions are captured from different views. Furthermore, these approaches are either not applicable or perform poorly when recognition is performed on videos from unknown and, more importantly, unseen views. Moreover, these methods do not scale well to new data and need to repeat the computationally expensive learning/training process when a new action class is to be added.

To simultaneously overcome these problems, we approach cross-view action recognition as a non-linear knowledge transfer learning problem where knowledge from multiple views is transferred to a single canonical view. Our approach consists of three phases. The first phase is unsupervised learning where a Non-linear Knowledge Transfer Model is learned. The proposed NKTM is a deep network with weight decay and sparsity constraints which finds a shared high-level virtual path that maps action videos captured from different viewpoints to the same canonical (i.e. frontal) view. The strongest point of our technique is that we learn a *single* NKTM for mapping all actions from all camera viewpoints to the same canonical view. Thus, action labels are not required while learning the NKTM or while transforming training and test actions to their respective canonical views using the NKTM. In the training phase, actions from unknown views are transformed to their corresponding canonical views using the learned NKTM. Action labels of training data are now required to train the subsequent classifier. In the final phase, actions from unknown and previously unseen views are transformed to their corresponding canonical views using the learned NKTM. The trained classifier is then used to classify the actions. We used a simple linear SVM to show the strength of the proposed NKTM. However, more sophisticated classifiers can also be used.

Our NKTM learning scheme is based on the observation that similar actions, when observed from different viewpoints, still have a common structure that puts them apart from other actions. Thus, it should be possible to separate action related features from viewpoint related features. The main challenge is that these features cannot be linearly separated. The second challenge comes from learning a non-linear model itself which requires large training data. Our solution is that we learn the NKTM from action trajectories of synthetic points fitted to mocap data. By projecting these points to different views, we can generate a large corpus of synthetic trajectories to learn the NKTM. We use k-means to generate a general codebook for encoding the action trajectories.

NKTM is learned once only from dense trajectories of synthetic points fitted to mocap data and then applied to real video data. Trajectories are coded with a general codebook learned from the same mocap data. NKTM is scalable to new action classes and training data as it does not require re-learning.

It is interesting to note that our technique outperforms the current cross-view action recognition methods on both IXMAS and N-UCLA datasets by transferring knowledge using the same NKTM learned without supervision (see Table 1). Therefore, compared to existing cross-view action recognition techniques, the proposed NKTM is more general and can be used in on-line action recognition systems. More precisely, the cost of adding a new action class using our NKTM in an on-line system is equal to training a multi-class SVM classifier. On the other hand, this situation is computationally expensive for most existing techniques [1, 4]. For instance nCTE [1] requires to perform computationally expensive spatio-temporal matching for each video sample of the new action class. Similarly, AOG [4] needs to train the AND/OR structure and tune its parameters. Compared to AOG [4] and nCTE [1], the training time of the proposed method is negligible. Thus, it can be used in an on-line system. Moreover, the test time of the proposed method is more faster than AOG [4] and comparable to nCTE. However, nCTE requires 30GB memory to store mocap samples.

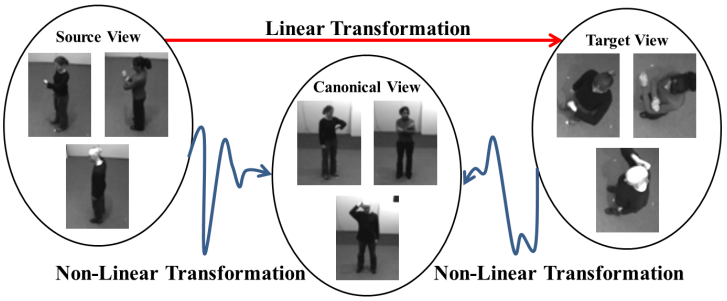

Figure 1: Existing cross-view action recognition techniques connect source and target views with a set of linear transformations that are unable to capture the non-linear manifolds on which real actions lie. Our NKTM finds a shared high-level non-linear virtual path that connects multiple source and target views to the same canonical view.

Table 1: Average recognition accuracy (%) on the IXMAS and N-UCLA Multiview datasets. DVV and CVP use samples from the target view. Our method neither requires target view samples nor joint positions.

	IXMAS	N-UCLA
Hankelets [2]	56.4	-
DVV [3]	38.2	51.0
CVP [5]	42.2	52.0
nCTE [1]	67.4	63.0
Proposed NKTM	**72.5**	**69.4**

[1] Ankur Gupta, Julieta Martinez, James J. Little, and Robert J. Woodham. 3D pose from motion for cross-view action recognition via non-linear circulant temporal encoding. In *CVPR*, 2014.

[2] Binlong Li, O.I. Camps, and M. Sznaier. Cross-view activity recognition using hankelets. In *CVPR*, 2012.

[3] R. Li and Todd Zickler. Discriminative virtual views for cross-view action recognition. In *CVPR*, 2012.

[4] J. Wang, X. Nie, Y. Xia, Y. Wu, and S.C. Zhu. Cross-view action modeling, learning and recognition. In *CVPR*, 2014.

[5] Zhong Zhang, Chunheng Wang, Baihua Xiao, Wen Zhou, Shuang Liu, and Cunzhao Shi. Cross-view action recognition via a continuous virtual path. In *CVPR*, 2013.

[6] Jingjing Zheng and Zhuolin Jiang. Learning view-invariant sparse representations for cross-view action recognition. In *ICCV*, 2013.

This is an extended abstract. The full paper is available at the Computer Vision Foundation webpage.