CROSS-VIEW ACTION RECOGNITION VIA LOW-RANK BASED DOMAIN ADAPTATION

Wen-Sheng Tseng[1,2], *Lun-Kai Hsu*[1,2], *Li-Wei Kang*[3], *and Yu-Chiang Frank Wang*[2]

[1] Dept. Electrical Engineering, National Taiwan University, Taipei, Taiwan
[2] Research Center for Information Technology Innovation, Academia Sinica, Taipei, Taiwan
[3] Dept. Comp. Sci. & Info. Eng., National Yunlin Univ. of Science & Technology, Yunlin, Taiwan

ABSTRACT

Cross-view action recognition is a challenging problem, since one typically does not have sufficient training data at the target view of interest. With recent developments of domain adaptation, we propose a novel low-rank based domain adaptation model for mapping labeled data from the original source view to the target view, so that training and testing can be performed at that domain. Our model not only provides an effective way for associating image data across different domains, we further advocate the structural incoherence between transformed data of different categories. As a result, additional data discriminating ability is introduced to our domain adaptation model, and thus improved recognition can be expected. Experimental results on the IXMAS dataset verify the effectiveness of our proposed method, which is shown to outperform state-of-the-art domain adaptation approaches.

Index Terms— Action recognition, domain adaptation, low-rank matrix decomposition

1. INTRODUCTION

Action recognition is among the active topics in computer vision and image processing. When recognizing actions in videos, action data are typically considered as spatiotemporal patterns, and existing works have proposed different visual features for representation [1, 2, 3, 4]. In practical scenarios, however, one needs to deal with action videos captured by different cameras (and thus at different views). Designing classifiers using training data at one camera view cannot be expected to generalize to recognize actions at a different view. As a result, recognizing actions across different views remains a very challenging task.

While geometry-based approaches have been proposed for cross-camera action recognition, [5, 6, 7], this type of methods typically require detection or tracking of body parts. Recently, techniques of *domain adaptation* have been utilized for solving this problem [8, 9]. The main idea of domain adaptation is to transfer the knowledge (e.g., feature or classifier) observed from one or few source domains to the target domain, so that the task in the target domain (e.g., recognizing actions captured by a new camera) can be solved accordingly. For example, Liu et al. [8] proposed to construct a bilingual codebook as a shared feature representation for action data captured by both domains. In [9], canonical correlation analysis was applied to derive the common feature space for cross-view action data. A SVM-based classifier taking the correlation information into consideration was designed for recognizing the projected data. While the above approaches have shown promising recognition results, they require the use of unlabeled data pairs which are obtained at both domains for deriving the feature space. In real-world scenarios for cross-view action recognition, collecting such cross-domain data pairs might not be applicable.

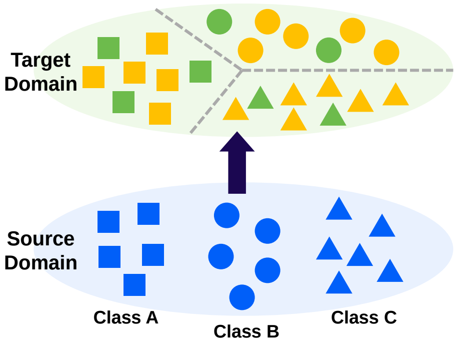

Fig. 1. Illustration of domain adaptation which projects source domain labeled data (colored in blue) into the target domain (i.e., those colored in orange). Together with the original small amount of training data at that domain (colored in green), recognition can be performed accordingly. Note that instances in different shapes indicate data of different classes.

In order to address the above practical problems, a novel domain adaptation approach was proposed in Jhuo et al. [10], who considered that data at the source domain can be reconstructed by a small amount of target domain data via a linear transformation. After mapping source domain data to the target domain, classifiers can be trained at that domain for recognition. Although they did not collect cross-domain data pairs as prior approaches did and reported promising results for cross-view image classification, they did not utilize label information during the process of domain adaptation.

With the practical setting in which source domain data can be obtained but only a small amount of labeled data are available at the target domain, we extend the idea of [10] and propose class-wise domain adaptation model with low-rank representation. By advocating the structural incoherence between the transformed data of different categories, we introduce additional discriminating ability into our proposed model. Once the process of domain adaptation is complete, we will be able to transform labeled source domain data into the target domain. Together with the training data at that domain, recognition can be performed accordingly. Later in experiments, we conduct experiments on a cross-view action recognition dataset. We will verify the effectiveness of our proposed method, which is shown to outperform state-of-the-art domain adaptation approaches.

2. LOW-RANK BASED DOMAIN ADAPTATION

2.1. Problem Formulation

As discussed in Section 1, we address the problem of cross-view action recognition under the scenarios in which labeled source view data are available, but only a small amount of training data are captured at the target view. Since there is no corresponding cross-view data pairs, methods like [8, 9] require sufficient cross-view data pairs cannot be easily applied for solving such domain adaptation problems.

In our work, we present a low-rank based domain adaptation approach for solving the above task. Suppose that we have source domain data $S = [S_1 ... S_N]$, where $S_i \in \mathbb{R}^{d\times n}$ indicates n instances of class i in a d-dimensional space, and N is the total number of classes. On the other hand, let $T = [T_1 ... T_N]$, in which $T_i \in \mathbb{R}^{d\times m}$ and $m < n$ (recall that only a small amount of training data at the target domain is available). In order to transfer source domain data into the target domain while preserving data discrimination, we propose to solve the following optimization problem:

$$\begin{aligned} \underset{W,Z_i,E_i}{min} \sum_{i=1}^{N} \{\|Z_i\|_* + \alpha\|E_i\|_{2,1}\} + \eta \sum_{i\neq j}^{N} \|Z_j^T Z_i\|_F^2 \\ s.t. \quad WS_i = TZ_i + E_i, WW^T = I, \end{aligned} \tag{1}$$

where W is observed for mapping source domain data into the target domain in terms of a linear combination of target domain data T. For each class i, $Z_i \in \mathbb{R}^{m\times n}$ and $E_i \in \mathbb{R}^{d\times n}$ are the weight and sparse error matrices, respectively. Inspired by low rank matrix decomposition [11], the minimization of the nuclear norm of each weight matrix Z_i aims at deriving a *compact* representation for better describing the transformed data in the target domain. The second term in (1) denotes the *structural incoherence* between the weights of different classes. Since this term promotes the incoherence between derived weight matrices of different classes, additional discriminating ability will be introduced into the derived model, and thus improved recognition performance can be expected.

It is worth noting that, we consider the transformed data WS_i of class i to be reconstructed by T instead of only T_i of the same class. This is because that, in most existing transfer learning and domain adaptation scenarios, the distributions of same-class data are expected to be very different across domains. Therefore, our proposed domain adaptation model in (1) does not require the transformed data to be represented by those of the associated class.

2.2. Optimization

To solve the optimization problem of (1), we first introduce an additional variable F and solve an equivalent problem as follows:

$$\begin{aligned} \underset{F_i,Z_i,E_i,W}{min} \sum_{i=1}^{N} \{\|F_i\|_* + \alpha\|E_i\|_{2,1}\} + \eta \sum_{i\neq j}^{N} \|F_j^T F_i\|_F^2 \\ s.t. \quad WS_i = TZ_i + E_i, Z_i = F_i. \end{aligned} \tag{2}$$

In stead of minimizing (2) directly, we solve the following class-wise optimization problem across different classes, until the derived domain adaptation model converges:

$$\begin{aligned} \underset{F_i,Z_i,E_i,W}{min} \|F_i\|_* + \alpha\|E_i\|_{2,1} + \eta \sum_{i\neq j}^{N} \|F_j^T F_i\|_F^2 \\ s.t. \quad WS_i = TZ_i + E_i, Z_i = F_i. \end{aligned} \tag{3}$$

It can be seen that, the above minimization problem involves the Frobenius norm of F_i and F_j pairs. To make this problem more tractable, we advance the property of $\|F_j^T F_i\|^2 \leq \|F_j\|^2\|F_i\|^2$ and relax (2) into the following formulation:

$$\begin{aligned} \underset{F_i,Z_i,E_i,W}{min} \|F_i\|_* + \alpha\|E_i\|_{2,1} + \eta'\|F_i\|_F^2 \\ s.t. \quad WS_i = TZ_i + E_i, Z_i = F_i, \end{aligned} \tag{4}$$

where $\eta' = \eta \sum_{j\neq i}^{N} \|F_j\|_F^2$ is a constant when solving the optimization problem for class i during each iteration. We now detail how we update the variable F, transformation W, weight matrix Z, and the resulting error matrix E for each class in the training stage of our method.

2.2.1. *Updating F_i*

We apply the technique of Augmented Lagrange Multiplier (ALM) [11] for solving the proposed optimization problem of (4). Given labeled source domain data of each class S_i and target domain training data T (of all classes), the Lagrangian function can be derived as follows:

$$\begin{aligned} &L(F_i, Z_i, E_i, W, X_i, Y_i) = \\ &\underset{F_i,Z_i,E_i,W,X_i,Y_i}{min} \eta'\|F_i\|_F^2 + \|F_i\|_* + \lambda\|E_i\|_{2,1} \\ &+ < X_i, Z_i - F_i > + \frac{\mu}{2}\|Z_i - F_i\|_F^2 \\ &+ < Y_i, WS_i - TZ_i - E_i > + \frac{\mu}{2}\|WS_i - TZ_i - E_i\|_F^2, \end{aligned} \tag{5}$$

where $< \cdot, \cdot >$ is the inner product operator. In (5), X_i and Y_i are the Lagrange multipliers, and $\mu > 0$ controls the convergence rate.

When updating F_i, we have Z_i, E_i, W, X_i, Y_i fixed in (5), and we calculate F_i^{k+1} in the $(k+1)$th iteration by:

$$\begin{aligned} F_i^{k+1} =& arg\,\underset{F_i}{min}\, \eta' \|F_i\|_F^2 + \|F_i\|_* + < X_i^k, Z_i^k - F_i > \\ & + \frac{\mu^k}{2} \|Z_i^k - F_i\|_F^2 \\ =& arg\,\underset{F_i}{min}\, \|F_i\|_* + (\eta' + \frac{\mu^k}{2}) < F_i, F_i > \\ & - \mu^k < Z_i^k + (\frac{1}{\mu^k} X_i^k), F_i > \\ =& arg\,\underset{F_i}{min} \epsilon\, \|F_i\|_* + \frac{1}{2} \|X_F - F_i\|_F^2, \end{aligned} \tag{6}$$

where $\epsilon = (2\eta' + \mu^k)^{-1}, X_F = \mu^k \epsilon (Z_i^k + \frac{1}{\mu^k} X_i^k)$. As suggested by [12], the solution of (6) can be solved as

$$F_i^{k+1} = U S_\epsilon V^T = U T_\epsilon [S] V^T, \tag{7}$$

where USV^T is the singular value decomposition of X_F. The operator $T_\epsilon[S]$ in (7) is defined by element-wise ϵ thresholding of S, i.e., $diag(T_\epsilon[S]) = [t_\epsilon[s_1], t_\epsilon[s_2], \ldots, t_\epsilon[s_r]]$ for rank(S) = r, and each $t_\epsilon[s]$ is determined as

$$t_\epsilon[s] = \begin{cases} s - \epsilon, & if\ s > \epsilon, \\ s + \epsilon, & if\ s < -\epsilon, \\ 0, & otherwise. \end{cases} \tag{8}$$

2.2.2. *Updating W*

To update W in the $(k+1)$th iteration, we fix the remaining variables and solve the following problem:

$$\begin{aligned} W^{k+1} =& arg\,\underset{W}{min} < Y_i^k, WS_i - TZ_i^k - E_i^k > \\ & + \frac{\mu^k}{2} \|WS_i - TZ_i^k - E_i^k\|_F^2. \end{aligned} \tag{9}$$

By taking the derivative of the above formulation to zero with respect to W, we have

$$Y_i^k S_i^T + \mu^k (WS_i - TZ_i^k - E_i^k) S_i^T = 0,$$

which yields the closed-form solution as

$$W^{k+1} = [(TZ_i^k + E_i^k) S_i^T - \frac{1}{\mu^k} Y_i^k S_i^T](S_i S_i^T)^{-1}. \tag{10}$$

2.2.3. *Updating Z_i*

We calculate Z_i in the $(k+1)$th iteration by fixing the remaining variables and solving the minimization problem below:

$$\begin{aligned} Z_i^{k+1} =& arg\,\underset{Z_i}{min} < X_i^k, Z_i - F_i^{k+1} > + \frac{\mu^k}{2} \|Z_i - F_i^{k+1}\|_F^2 \\ & + < Y_i^k, W^{k+1} S_i - TZ_i - E_i^k > \\ & + \frac{\mu^k}{2} \|W^{k+1} S_i - TZ_i - E_i^k\|_F^2. \end{aligned} \tag{11}$$

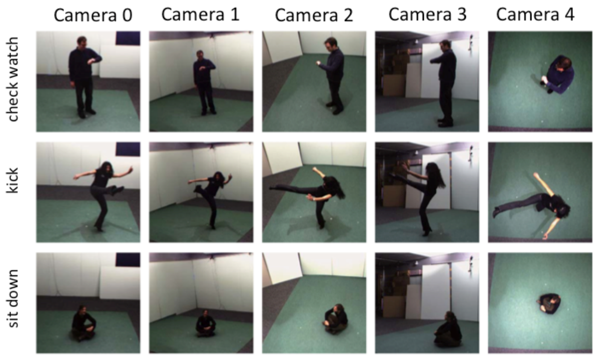

Fig. 2. Example actions of the IXMAS dataset. Each row represents an action at five different views.

Similar to the case of updating W, we set the derivative of the above formulation to zero with respect to Z_i. As a result, we have the closed-form solution of Z_i as

$$\begin{aligned} Z_i^{k+1} =& (I + T^T T)^{-1} [T^T (W^{k+1} S_i - E_i^k) + \\ & \frac{1}{\mu^k} (T^T Y_i^k - X_i^k) + F_i^{k+1}]. \end{aligned} \tag{12}$$

2.2.4. *Updating E_i*

Finally, we update E_i by fixing other variables and solve:

$$\begin{aligned} E_i^{k+1} =& arg\,\underset{E_i}{min}\, \lambda \|E_i\|_{2,1} + < Y_i^k, W^{k+1} S_i - TZ_i^{k+1} - E_i > \\ & + \frac{\mu^k}{2} \|W^{k+1} S_i - TZ_i^{k+1} - E_i\|_F^2 \\ =& arg\,\underset{E_i}{min}\, \epsilon' \|E_i\|_{2,1} + \frac{1}{2} \|E_i - X_e\|_F^2, \end{aligned} \tag{13}$$

where $\epsilon' = (\lambda / \mu^k)$ and $X_e = W^{k+1} S_i - TZ_i^{k+1} + \frac{Y_i^k}{\mu^k}$. The above optimization problem can be solved by ℓ_1-minimization techniques such as [13].

2.3. Performing Recognition

Once the transformation W and the associated weight matrices Z_i for each class are observed, the domain adaptation process is complete. To train classifiers for recognizing target view data, we first map the source view labeled data S_i of each class into the target domain using W. Together with target view training data T_i, we train SVM classifiers at the target domain for recognizing target view test data.

3. EXPERIMENTAL RESULTS

To evaluate the performance of cross-view action recognition, we consider the IXMAS dataset [7] for experiment. This dataset contains the action videos of eleven classes. Each action is performed by twelve actors for three times, and is captured by five different cameras (see examples shown in Figure 2). We extract descriptors defined by [15] and describe

Table 1. Performance comparisons on the IXMAS dataset. Note that each column indicates the source view camera (for training), and each row is the target view camera for recognition. The seven approaches denoted by A to G are: (A) SVM trained by target domain images only, (B) SVM trained by source domain images only, (C) direct combination of source and target domain images for training SVM, (D) CCA [14], (E) BoBW [8], (F) RDALR [10], and (G) our proposed method. Note that methods of A and B do not perform domain adaptation.

	camera 0							camera 1						
	A	B	C	D	E	F	G	A	B	C	D	E	F	G
cam 0				-				0.48	9.63	16.58	15.90	12.63	**37.17**	**37.17**
cam 1	0.05	11.02	21.82	17.17	22.22	16.04	**25.60**				-			
cam 2	2.41	10.27	18.29	15.40	2.78	6.68	**25.67**	1.18	9.73	12.09	12.88	5.81	24.60	**27.27**
cam 3	0.32	10.16	17.43	10.35	8.84	10.16	**27.54**	0.53	9.73	14.39	13.39	5.05	17.65	**27.27**
cam 4	2.03	11.71	17.70	11.87	18.94	26.47	**26.73**	1.44	8.50	12.46	12.88	4.29	**30.48**	**30.48**
Avg.	1.20	10.79	18.81	13.70	13.19	14.84	**26.13**	0.91	9.40	13.88	13.76	6.94	27.47	**30.55**

	camera 2							camera 3						
	A	B	C	D	E	F	G	A	B	C	D	E	F	G
cam 0	0.64	12.14	33.52	22.22	26.26	**37.97**	30.75	2.30	13.48	**35.13**	6.57	4.29	21.93	33.96
cam 1	0.64	13.26	**34.28**	19.70	4.29	32.36	28.88	0.96	13.53	**33.53**	18.94	16.16	11.23	31.55
cam 2				-				1.55	6.42	**30.59**	16.92	19.19	5.88	26.20
cam 3	0.27	7.27	22.14	14.39	28.28	9.36	**29.95**				-			
cam 4	0.91	10.43	21.39	12.63	17.68	23.53	**26.47**	0.91	9.52	25.83	15.40	15.40	21.93	**25.94**
Avg.	0.61	10.78	27.78	17.23	19.13	25.80	**29.01**	1.43	10.74	**31.27**	14.46	13.76	15.24	29.41

each action video as a group of spatio-temporal cuboids (at most 200). For each view these cuboids are quantized into N = 1000 visual words. In our experiments, we choose one camera view as the source domain and one of another camera views to be recognized as the target domain. For each class, we use all 36 instances in the source domain and randomly select 2 instances from the target domain for learning our domain adaptation model. Once this learning stage is complete, we collect the all 36 instances (projected from the source domain) and the 2 instances at the target domain for training SVM classifiers. The remaining 34 instances of each class at the target view will be the test data.

We compare our method with several baseline or state-of-the-art approaches: (A) training using only target view images (2 per class), (B) training using source view labeled images only (36 per class), (C) direct training of source and target view images (36+2 images per class), (D) CCA (canonical correlation analysis)-based approach [14], (E) BoBW of [8], and (F) RDALR (robust domain adaptation with low-rank reconstruction) [10]. Both CCA and BoBW require corresponding source and target-view data pairs for deriving the joint feature representation. For fair comparisons, we use the same numbers of training images for both CCA and BoBW (i.e., two pairs of source-target view images for deriving their domain adaptation models, and 34 labeled source view images to be transformed into the target domain). We note that, we do not consider the setting of transferring top view data (camera 4) into other views in our experiments, since it is not expected to generalize well for all approaches.

Table 1 lists the average recognition results for difference approaches. It can be seen that direct training using target domain data (i.e., column A) produced very poor recognition performance due to insufficient data. As shown in column B, using labeled source domain data for training classifiers did not provide generalization due to domain changes. While a naive combination of training data of A and B produced much better recognition results, domain adaptation methods (D to G) generally improved the performance. Among different approaches, we see that our method achieved the best or comparable results in Table 1, and thus this verifies the effectiveness of our proposed domain adaptation approach.

4. CONCLUSIONS

In this paper, we presented a low-rank based domain adaptation model for solving cross-view action recognition problems. Our proposed model aims at projecting source domain data into the target domain via a low-rank based representation, which better describes the transformed data at the target domain in a compact and representative way. By advocating the structural incoherence between the observed representations of different classes, we introduced additional discriminating ability into our proposed model, and thus our model can be applied for addressing cross-domain classification problems. From experimental results on the IXMAS dataset, we confirmed that our method outperformed baseline and state-of-the-art domain adaptation approaches on cross-view action recognition. Thus, the effectiveness of our proposed model can be successfully verified.

Acknowledgement This work is supported in part by National Science Council of Taiwan via NSC100-2221-E-001-018-MY2 and NSC101-2631-H-001-007.

References

[1] G. Mori, X. Ren, A. A. Efros, and J. Malik, "Recovering human body configurations: combining segmentation and recognition," *IEEE CVPR*, 2004.

[2] M. Blank, L. Gorelick, E. Shechtman, M. Irani, and R. Basri, "Actions as space-time shapes," *IEEE ICCV*, 2005.

[3] Z. Lin, Z. Jiang, and L. S. Davis, "Recognizing actions by shapemotion prototype trees," *IEEE ICCV*, 2009.

[4] J. Liu, Y. Yang, and M. Shah, "Learning semantic visual vocabuaries using diffusion distance," *IEEE CVPR*, 2009.

[5] C. Rao, A. Yilmaz, and M. Shah, "View-invariant representation and recognition of actions," *IJCV*, 2002.

[6] V. Paramesmaran and R. Chellappa, "View invariance for human action recognition," *IJCV*, 2006.

[7] D. Weinland, M. Ozuysal, and P. Fua, "Making action recognition robust to occlusions and viewpoint changes," *ECCV*, 2010.

[8] J. Liu, M. Shah, B. Kuipers, and S. Savarese, "Cross-view action recognition via view knowledge transfer," *IEEE CVPR*, 2011.

[9] C.-H. Huang, Y.-R. Yeh, and Y.-C. F. Wang, "Recognizing actions across cameras by exploring the correlated subspace," *ECCV Workshop on Video Event Categorization, Tagging and Retrieval*, 2012.

[10] I.-H. Jhuo, D. Liu, D. T. Lee, and S. F. Chang, "Robust visual domain adaptation with low-rank reconstruction," *IEEE CVPR*, 2012.

[11] E. Candès, X. Li, Y. Ma, and J. Wright, "Robust principal component analysis?," *Journal of ACM*, 2011.

[12] J. Cai, E. Candès, and Z. Shen, "A singular value thresholding algorithm for matrix completion," *SIAM Jour. Optimization*, 2010.

[13] Elaine T. Hale, Wotao Yin, and Yin Zhang, "Fixed-point continuation for ℓ_1-minimization: Methodology and convergence," *SIAM Jour. Optimization*, 2008.

[14] H. Hotelling, "Relations between two sets of variates," *Biometrika*, vol. 28, no. 3/4, pp. 321–377, Dec. 1936.

[15] P. Dollar, V. Rabaud, G. Cottrell, and S. Belongie, "Behavior recognition via sparse spatio-temporal features," *ICCV Workshop on VS-PETS*, 2005.