LETTER

Robust Transferable Subspace Learning for Cross-Corpus Facial Expression Recognition*

Dongliang CHEN[†], ***Nonmember***, **Peng SONG**[†a)], ***Member***, **Wenjing ZHANG**[†], **Weijian ZHANG**[†], **Bingui XU**[††], ***and*** **Xuan ZHOU**[††], ***Nonmembers***

SUMMARY In this letter, we propose a novel robust transferable subspace learning (RTSL) method for cross-corpus facial expression recognition. In this method, on one hand, we present a novel distance metric algorithm, which jointly considers the local and global distance distribution measure, to reduce the cross-corpus mismatch. On the other hand, we design a label guidance strategy to improve the discriminate ability of subspace. Thus, the RTSL is much more robust to the cross-corpus recognition problem than traditional transfer learning methods. We conduct extensive experiments on several facial expression corpora to evaluate the recognition performance of RTSL. The results demonstrate the superiority of the proposed method over some state-of-the-art methods.

key words: *facial expression recognition, subspace learning, transfer learning, graph Laplacian*

1. Introduction

Facial expression recognition has become an active research topic because of its far-reaching applications in human-computer interaction, multimedia entertainment, machine intelligence, medicine and psychology [1]. The main purpose of facial expression recognition is to recognize the unlabeled facial images into various emotional states, e.g., anger, disgust, fear, happiness, sadness, and surprise.

Current facial expression recognition methods can achieve satisfactory performance under restricted conditions. However, in practice, the training and testing data are often sampled from different corpora, which are recorded from different devices or environments. This would lead to large feature distribution divergence, and suffer a heavy drop in performance. Thus, it is worthwhile to investigate the cross-corpus facial expression recognition problem.

To address the above-mentioned challenging problem, over the past few years, with the development of transfer learning [2], many methods have been developed. For example, in [3], Chu et al. propose a simple yet effective transfer learning method called selective transfer machine (STM), STM can simultaneously learn a classifier and re-weight the training samples that the most relevant to the testing subject. In [4], Zheng et al. have presented a transductive transfer regularized least-squares regression (TTRLSR) model to cope with the cross-domain color facial expression recognition problem. In [5], Yan et al. develop an unsupervised domain adaptive dictionary learning (UDADL) method, which aims to learn a shared dictionary to bridge the source and target samples. However, most of existing methods only consider a global distance metric, i.e., MMD, in which the learned corpus-invariant features may not only draw both datasets close, but also mix all the data points with different classes togethe. Thus, the discriminative ability of the common subspace is vital for cross-corpus recognition [6].

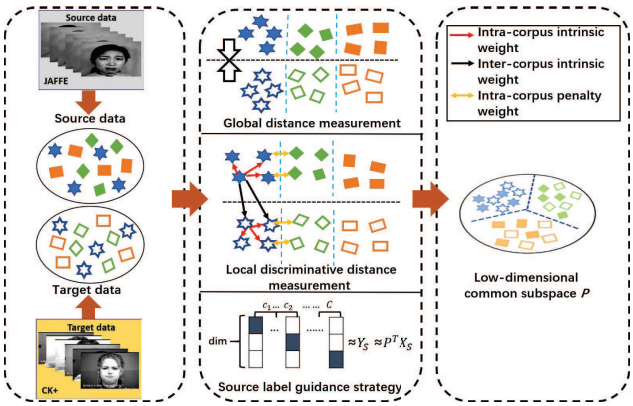


Fig. 1 The diagram of our proposed RTSL method.

In this letter, to cope with the cross-corpus facial expression recognition problem, we propose a novel robust transferable subspace learning (RTSL) method. Different from the above-mentioned methods, our method not only can reduce the feature distribution divergence by utilizing a novel distance metric, but also can capture the discriminative knowledge of cross-corpus sample. Therefore, our method can learn a low-dimensional common feature subspace for source and target corpora, and it is much more robust to the cross-corpus recognition problem than traditional transfer learning methods. Figure 1 shows the diagram of our approach.

2. Proposed Method

To begin with, we briefly introduce some notations fre-

Manuscript received May 21, 2020.
Manuscript revised July 10, 2020.
Manuscript publicized July 20, 2020.

[†]The authors are with the School of Computer and Control Engineering, Yantai University, Yantai 264005, P.R. China.
[††]The authors are with the Shandong Institute of Space Electronic Technology, Yantai 264670, P.R. China.

*The work is partly supported by the National Natural Science Foundation of China under Grant 61703360 and the Fundamental Research Funds for the Central Universities under Grant CDLS-2019-01.

a) E-mail: pengsong@ytu.edu.cn (Corresponding author)

DOI: 10.1587/transinf.2020EDL8074

quently used in this letter. We denote $X_S \in R^{m\times N_S}$ be the emotional features of source corpus, and $Y_S = \{y_{S\,i}\}_{i=1}^{N_S}$ be the corresponding labels. Similarly, let $X_T \in R^{m\times N_T}$ be the emotional features of target corpus. N_S and N_T are the corresponding numbers of source and target samples, respectively, and m is the dimension of feature vectors. Then, we introduce a feature matrix $X = [X_S, X_T] = \{x_i\}_{i=1}^{N} \in R^{m\times N}$ as the input matrix, where $N = N_S + N_T$.

In this work, we devote to address the cross-corpus facial expression recognition problem. Thus, we expect that our method can efficiently reduce the feature distribution divergence and obtain robust corpus-invariant feature representations. To achieve this goal, we first introduce a global distance metric algorithm for source and target data, in which both marginal MMD [7] and conditional MMD [8] are considered to measure the feature distribution divergence, which can be formulated as

$$\min_P Tr(P^T X(M_0 + M_C)X^T P) + \gamma\|P\|_F^2 \tag{1}$$

where P is the projection matrix, $\gamma > 0$ is a trade-off parameter, M_0 and $M_C = \sum_{c=1}^{C} M_c$ are the marginal and conditional MMD matrices, respectively. In detail, the marginal MMD can be computed as

$$Tr(P^T X M_0 X^T P) = \left\| \frac{1}{N_S}\sum_{i=1}^{N_S} P^T x_i - \frac{1}{N_T}\sum_{j=N_S+1}^{N} P^T x_j \right\|^2 \tag{2}$$

and the conditional MMD can be calculated by

$$Tr(P^T X M_C X^T P) = \sum_{c=1}^{C} \left\| \frac{1}{N_S^c}\sum_{x_i\in\mathcal{D}_S^{(c)}} P^T x_i - \frac{1}{N_T^c}\sum_{x_j\in\mathcal{D}_T^{(c)}} P^T x_j \right\|^2 \tag{3}$$

where C is the number of classes, $\mathcal{D}_S^{(c)}$ denotes the set of source samples with their true class labels belonging to class c, and $\mathcal{D}_T^{(c)}$ denotes the set of target samples with their pseudo class labels belonging to class c. Specifically, the labels of target data are unavailable in training, thus we utilize the target pseudo labels to compute M_0.

By using the global distance metric in (1), we can reduce the feature distribution divergence between source and target corpora. However, it does not take into account the discriminative knowledge and local manifold structures of sample points, which has been proven very useful for feature representation [9]. Thus, we present a novel local discriminative distance metric by aligning the geometric structure with label information, which is formulated as

$$\min_P Tr(P^T X(L_w - \lambda L_b)X^T P) = \min_P Tr(P^T X L X^T P) \tag{4}$$

where L_w and L_b are Laplacian matrices of dual intrinsic graph and total penalty graph for source and target data, respectively, $\lambda > 0$ is a trade-off parameter. Mathematically, $L_w = D_w - W_w$, $L_b = D_b - W_b$, where D_w and D_b are the diagonal matrices, W_w and W_b are the weight matrices for the dual intrinsic graph and the total penalty graph, respectively. In this work, we deploy the following two criteria to construct W_w and W_b:

1) *Construct the dual intrinsic weight matrix W_w*: For cross-corpus data, we expect that our method can minimize the intra-class compactness, meanwhile, it can reduce the feature divergence between source and target corpora. To this end, we design a dual intrinsic weight matrix W_w by considering both intra-corpus and inter-corpus similarities:

- Intra-corpus intrinsic weight matrix W_w^S and W_w^T:

$$(W_w^S)_{ij} = \begin{cases} 1, & if\ i \neq j, y_{S_i} = y_{S_j} \\ 0, & otherwise \end{cases} \tag{5}$$

$$(W_w^T)_{ij} = \begin{cases} 1, & if\ i \neq j, \hat{y_T}_i = \hat{y_T}_j \\ 0, & otherwise \end{cases} \tag{6}$$

  where $\hat{y_T}$ is the pseudo labels of target data.
- Inter-corpus intrinsic weight matrix W_w^{ST} and W_w^{TS}:

$$(W_w^{ST})_{ij} = \begin{cases} 1, & x_j \in N_k^c(x_i) or x_i \in N_k^c(x_j) \\ 0, & otherwise \end{cases} \tag{7}$$

$$(W_w^{TS})_{ji} = \begin{cases} 1, & x_i \in N_k^c(x_j) or x_j \in N_k^c(x_i) \\ 0, & otherwise \end{cases} \tag{8}$$

  where x_i and x_j are from different datasets, and $N_k^c(x_i)$ indicates the index set of the k_1-nearest neighbors of x_i in the same class.

By combining these two kinds of similarity weight matrices, we can construct the dual intrinsic weight matrix W_w as follows:

$$W_w = \begin{bmatrix} W^S & W^{ST} \\ W^{TS} & W^T \end{bmatrix} \tag{9}$$

2) *Construct the total penalty weight matrix W_b*: To explore more class-discriminative information, we attempt to maximize the inter-class separability. Thus, we construct two intra-corpus penalty weight matrices W_b^S and W_b^T, defined by

$$(W_b^S)_{ij}, (W_b^T)_{ij} = \begin{cases} 1, & x_j \in P_k^d(x_i) or x_i \in P_k^d(x_j) \\ 0, & otherwise \end{cases} \tag{10}$$

where $P_k^d(x_i)$ indicates the index set of the k_2-nearest neighbors of x_i in distinct classes. By combining W_b^S and W_b^T, we can obtain the total penalty weight matrix W_b as

$$W_b = \begin{bmatrix} W^S & 0 \\ 0 & W^T \end{bmatrix}. \tag{11}$$

To further improve the discriminative ability of the learned low-dimensional common subspace, we implement a source label guidance strategy, in which a linear regression function is adopted:

$$\min_{P,V} \frac{1}{2}\left\|P^T X_S - (Y_S + B \odot V)\right\|_F^2 \quad s.t.\ V \geq 0 \tag{12}$$

where V is a non-negative label relaxation matrix, B is a luxury matrix, $\odot$ indicates a Hadamard product operator of matrices, and Y_S is the label matrix of the source samples. In detail, Y_S and B are defined as

$$Y_{S\{i,j\}} = \begin{cases} 1, & if\ y_{S_j} \in \text{the i}-\text{th class} \\ 0, & otherwise \end{cases} \tag{13}$$

$$B_{i,j} = \begin{cases} +1, & if\ Y_{S\{i,j\}} = 1 \\ -1, & if\ Y_{S\{i,j\}} = 0 \end{cases} \tag{14}$$

where $i = 1, \ldots, C$ and $j = 1, \ldots, N_S$.

By combining Eqs. (1), (4) and (12), we can obtain the objective function of RTSL as

$$\begin{aligned} \min_{P,V}\ & Tr(P^T X(\alpha M + \beta L)X^T P) + \gamma \|P\|_F^2 \\ & + \frac{1}{2}\left\|P^T X_S - (Y_S + B \odot V)\right\|_F^2 \\ s.t.\ & P^T P = I, V \geq 0 \end{aligned} \tag{15}$$

where $\alpha > 0$ and $\beta > 0$ are the trade-off parameters, $L = L_w - \lambda L_b$, and $M = M_0 + M_C$.

To solve the objective function in (15), we present an iterative optimization algorithm, and Eq. (15) can be reformulated as

$$\begin{aligned} \mathcal{L} = & Tr(P^T X(\alpha M + \beta L)X^T P) + \gamma \|P\|_F^2 + Tr(\phi(I - P^T P)) \\ & + \frac{1}{2}\left\|P^T X_S - (Y_S + B \odot V)\right\|_F^2 \end{aligned} \tag{16}$$

where ϕ is a Lagrange parameter. Then the problem (16) can be optimized by an iterative manner, which is given as

1) **Update** P: Fix V, we can update P by minimizing the problem (16). By setting the derivative $\partial \mathcal{L}/\partial P = 0$, we can obtain the variable P^* as

$$P^* = (2XG_1X^T + 2G_2 + X_S X_S^T)^{-1}(X_S(Y_S + B \odot V)) \tag{17}$$

where $G_1 = (\alpha M + \beta L)$ and $G_2 = (\gamma - \phi)I$.

2) **Update** V: Fix P, we can update V by minimizing the following problem

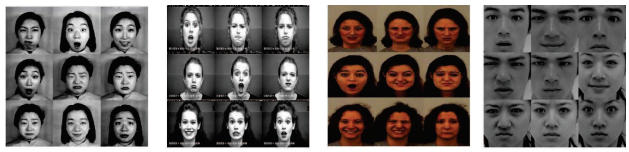
$$\min_{V \geq 0} \frac{1}{2}\left\|P^T X_S - (Y_S + B \odot V)\right\|_F^2 \tag{18}$$

According to Ref. [10], the optimal solution of V^* can be rewritten as $V^* = \max\left\{(P^T X_S - Y_S) \odot B, 0\right\}$.

It is worth noting that, at the beginning, some of the target pseudo labels may be incorrect. Thus, we employ an iterative manner to update the labels with the progressive learning of common subspace, which can alternatively improve the labeling quality until convergence. Specially, we utilize a simple SVM classifier to obtain target pseudo labels in each iteration. Also, it should be noted that the dimension of the learned common subspace is equal to the number of emotion category.

3. Experiments

To evaluate the performance of our method, we conduct extensive experiments of cross-corpus facial expression recognition on four publicly available facial expression datasets, including JAFFE[†] [11], CK+[††] [12], KDEF[†††] [13] and TFEID[††††] [14]. Figure 2 shows the examples of these datasets.



(a) (b) (c) (d)

Fig. 2 Examples of facial images with different expressions from (a) JAFFE, (b) CK+, (c) KDEF and (d) TFEID.

We select six common basic expressions of these datasets, i.e., *Anger, Disgust, Fear, Happiness, Sadness, and Surprise*. Then, we crop and transform these facial images to the size 60×60 and extract LBP features. Specifically, we divide each facial image into 9 (3 × 3) regions and use a 2304 (256×9) dimensional LBP feature accordingly. Based on these datasets, we conduct 12 different settings of experiments for cross-corpus recognition (source → target), i.e., $J \rightarrow C$, $J \rightarrow K$, $J \rightarrow T$, $C \rightarrow J$, $C \rightarrow K$, $C \rightarrow T$, $K \rightarrow J$, $K \rightarrow C$, $K \rightarrow T$, $T \rightarrow J$, $T \rightarrow C$, and $T \rightarrow K$, where *J, C, K* and *T* are short for JAFFE, CK+, KDEF and TFEID.

In our experiments, we compare our method with recently related state-of-the-art methods including geodesic flow kernel (GFK) [15], transfer component analysis (TCA) [7], joint distribution adaptation (JDA) [8], transfer joint matching (TJM) [16], discriminative transfer subspace learning (DTSL) [10], domain invariant and class discriminative feature learning (DICD) [6], and principal component analysis (PCA). Then, we choose SVM as the baseline classifier, in which the classifier trained on labeled source data is adopted to classify the unlabeled target data. For SVM, all the parameters (i.e., penalty term, bandwidth of RBF kernel σ) are chosen by a grid-search strategy.

Since the source and target data follow different feature distributions, we cannot automatically select the optimal model parameters under the cross-validation strategy [8]. Therefore, we empirically search the parameter space for the optimal values to evaluate and report the best results of each method. For the parameters of RTSL, we set the number of nearest neighbors $k_1 = 5$, $k_2 = 20$, α and γ are tuned from the parameter set [0.1,1,10], and β and λ are tuned from the parameter set [0.01,0.1,1]. For all the baseline methods, we report results in the original paper or the best we can get. For the PCA, TCA, JDA, TJM, and DICD, the subspace dimension is set to 100. The subspace dimension of DTSL and RTSL is set to 6, in which the subspace dimension of these two algorithms is equal to the number of classes [10]. Finally, we use the classification accuracy on the testing target corpus to measure the performance.

Table 1 shows the recognition results of our proposed

[†] http://www.kasrl.org/jaffe.html
[††] http://www.pitt.edu/~emotion/ck-spread.htm
[†††] http://www.emotionlab.se/kdef/
[††††] http://bml.ym.edu.tw/tfeid/

Table 1 Recognition accuracy (%) of different methods under different settings.

Tasks	Compared methods										
	PCA	GFK	TCA	JDA	TJM	DTSL	DICD	RTSL	$RTSL_a$	$RTSL_b$	$RTSL_c$
$J \rightarrow C$	39.52	31.90	41.90	43.33	40.48	49.52	46.67	**50.48**	50.00	48.10	45.71
$J \rightarrow K$	41.90	34.76	48.10	51.43	42.38	47.14	50.95	54.29	**54.76**	50.48	42.38
$J \rightarrow T$	20.95	23.33	39.05	39.52	35.71	44.29	41.43	**44.76**	42.38	41.43	36.67
$C \rightarrow J$	33.33	34.43	39.89	42.62	40.98	38.80	44.81	**45.36**	43.17	42.62	37.70
$C \rightarrow K$	46.67	42.85	50.48	50.95	47.62	51.43	58.57	60.48	59.05	56.67	**61.43**
$C \rightarrow T$	34.76	45.71	44.76	43.33	42.38	44.76	**47.62**	**47.62**	47.14	46.19	40.48
$K \rightarrow J$	39.34	40.98	44.81	45.90	49.18	48.09	46.45	**50.27**	48.63	49.18	43.16
$K \rightarrow C$	52.38	56.67	53.81	55.71	53.33	61.43	61.90	**66.19**	63.33	62.38	64.76
$K \rightarrow T$	40.48	43.33	44.76	46.67	45.71	50.00	49.52	**50.47**	**50.47**	47.14	40.00
$T \rightarrow J$	22.95	25.13	41.53	40.44	40.98	39.89	42.62	**45.36**	42.62	44.26	40.98
$T \rightarrow C$	40.00	36.67	50.95	52.86	43.81	53.33	51.90	**56.67**	52.86	53.81	50.48
$T \rightarrow K$	40.95	43.81	48.57	48.10	43.33	49.52	50.48	**55.24**	54.76	54.29	47.62
Average	37.77	38.30	45.72	46.74	43.82	48.18	49.41	**52.27**	50.76	49.71	45.94

RTSL method and seven baseline methods. From the table, we can have the following observations. First, among all the transfer learning algorithms, our RTSL method achieves the best recognition performance in all cases. This reasons might be two-fold. On one hand, RTSL utilizes a novel distance metric algorithm to jointly reduce the distribution divergence, in which the global and local distance measurement are considered together. Specifically, the former aims to align the global feature distribution of two corpora, and the latter considers to align the similar source and target samples. On the other hand, the discriminative information is considered in our transfer learning framework. Second, most of transfer learning methods including RTSL achieve better performance than the traditional PCA algorithm. This can be attributed to the power of transfer learning. Third, it is interesting to find that the recognition rates on the two cases, i.e., $C \rightarrow K$ and $K \rightarrow C$, are much higher that those on the other cases. The reason might be that the expression styles are similar on CK+ and KDEF datasets. Finally, compared with DICD, which also aims to learn a domain-invariant and class-invariant feature representation, our RTSL method can significantly achieve higher recognition accuracies.

To further verify the effectiveness of our method, we consider three special cases of RTSL, i.e., $RTSL_a$ (neglecting the source label guidance), $RTSL_b$ (neglecting the local distance measurement, $\beta = 0$) and $RTSL_c$ (neglecting the global distance measurement, $\alpha = 0$). Note that since $RTSL_a$ does not consider the source label guidance, following the experimental settings of subspace learning algorithms, the subspace dimension in $RTSL_a$ is set to 100, while as RTSL, the subspace dimension in $RTSL_b$ and $RTSL_c$ is 6. The results are given in Table 1. From the table, we can find that the novel global and local distance measurement plays an important role in our model. In particular, the global distance measurement has the largest influence to our model. In addition, the source label guidance also improves the recognition performance of RTSL. These results verify that RTSL is effective for cross-corpus facial expression recognition. Moreover, it is surprising to find that, in J $\rightarrow$ K and K $\rightarrow$ T, $RTSL_a$ achieves the highest recognition performance. This indicates that the distance measurement might be much more important than the source label guidance in some cases. In C $\rightarrow$ K, $RTSL_c$ performs the best, and in K $\rightarrow$ C, $RTSL_c$ also outperforms $RTSL_b$. These results indicate that, in some cases, the local discriminative distance measurement might play a much more effective role than the global distance measurement.

4. Conclusion

In this letter, we have presented a novel transfer learning method, called robust transferable subspace learning (RTSL), to cope with the cross-corpus facial expression recognition problem. The main contribution of RTSL lies in that it can effectively reduce the feature distribution divergence, and obtain the robust corpus-invariant feature representations for source and target data. Experimental results on several benchmarks verify that our method can significantly outperform some state-of-the-art transfer learning methods.

References

[1] Z. Zeng, M. Pantic, G.I. Roisman, and T.S. Huang, "A survey of affect recognition methods: Audio, visual, and spontaneous expressions," IEEE Trans. Pattern Anal. Mach. Intell., vol.31, no.1, pp.39–58, 2008.

[2] J. Zhang, W. Li, P. Ogunbona, and D. Xu, "Recent advances in transfer learning for cross-dataset visual recognition: A problem-oriented perspective," ACM Comput. Surv., vol.52, no.1, pp.1–38, 2019.

[3] W.-S. Chu, F.D. Torre, and J.F. Cohn, "Selective transfer machine for personalized facial expression analysis," IEEE Trans. Pattern Anal. Mach. Intell., vol.39, no.3, pp.529–545, 2016.
[4] W. Zheng, Y. Zong, X. Zhou, and M. Xin, "Cross-domain color facial expression recognition using transductive transfer subspace learning," IEEE Trans. Affective Comput., vol.9, no.1, pp.21–37, 2016.
[5] K. Yan, W. Zheng, Z. Cui, Y. Zong, T. Zhang, and C. Tang, "Unsupervised facial expression recognition using domain adaptation based dictionary learning approach," Neurocomputing, vol.319, pp.84–91, 2018.
[6] S. Li, S. Song, G. Huang, Z. Ding, and C. Wu, "Domain invariant and class discriminative feature learning for visual domain adaptation," IEEE Trans. Image Process., vol.27, no.9, pp.4260–4273, 2018.
[7] S.J. Pan, I.W. Tsang, J.T. Kwok, and Q. Yang, "Domain adaptation via transfer component analysis," IEEE Trans. Neural Netw., vol.22, no.2, pp.199–210, 2011.
[8] M. Long, J. Wang, G. Ding, J. Sun, and P.S. Yu, "Transfer feature learning with joint distribution adaptation," Proc. IEEE International Conference on Computer Vision, pp.2200–2207, 2013.
[9] P. Song, "Transfer linear subspace learning for cross-corpus speech emotion recognition," IEEE Trans. Affective Comput., vol.10, no.2, pp.265–275, 2019.
[10] Y. Xu, X. Fang, J. Wu, X. Li, and D. Zhang, "Discriminative transfer subspace learning via low-rank and sparse representation," IEEE Trans. Image Process., vol.25, no.2, pp.850–863, 2015.
[11] M.J. Lyons, J. Budynek, and S. Akamatsu, "Automatic classification of single facial images," IEEE Trans. Pattern Anal. Mach. Intell., vol.21, no.12, pp.1357–1362, 1999.
[12] P. Lucey, J.F. Cohn, T. Kanade, J. Saragih, Z. Ambadar, and I. Matthews, "The extended cohn-kanade dataset (ck+): A complete dataset for action unit and emotion-specified expression," 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition - Workshops, pp.94–101, 2010.
[13] D. Lundqvist, A. Flykt, and A. Öhman, "The karolinska directed emotional faces (kdef)," CD ROM from Department of Clinical Neuroscience, Psychology section, Karolinska Institutet, vol.91, p.630, 1998.
[14] L.F. Chen and Y.S. Yen, "Taiwanese facial expression image database," Brain Mapping Laboratory, Institute of Brain Science, National Yang-Ming University, Taipei, Taiwan, 2007.
[15] B. Gong, Y. Shi, F. Sha, and K. Grauman, "Geodesic flow kernel for unsupervised domain adaptation," 2012 IEEE Conference on Computer Vision and Pattern Recognition, pp.2066–2073, 2012.
[16] M. Long, J. Wang, G. Ding, J. Sun, and P.S. Yu, "Transfer joint matching for unsupervised domain adaptation," Proc. IEEE Conference on Computer Vision and Pattern Recognition, pp.1410–1417, 2014.