Robust Multi-view Representation: A Unified Perspective from Multi-view Learning to Domain Adaption

Zhengming Ding[†], **Ming Shao**[‡], **Yun Fu**[†♯]
[†]Department of ECE, College of Engineering, Northeastern University, Boston, USA
[‡]Computer and Information Science, University of Massachusetts Dartmouth, USA
[♯]College of Computer and Information Science, Northeastern University, Boston, USA
allanding@ece.neu.edu, mshao@umassd.edu, yunfu@ece.neu.edu

Abstract

Multi-view data are extensively accessible nowadays thanks to various types of features, different view-points and sensors which tend to facilitate better representation in many key applications. This survey covers the topic of robust multi-view data representation, centered around several major visual applications. First of all, we formulate a unified learning framework which is able to model most existing multi-view learning and domain adaptation in this line. Following this, we conduct a comprehensive discussion across these two problems by reviewing the algorithms along these two topics, including multi-view clustering, multi-view classification, zero-shot learning, and domain adaption. We further present more practical challenges in multi-view data analysis. Finally, we discuss future research including incomplete, unbalance, large-scale multi-view learning. This would benefit AI community from literature review to future direction.

1 Introduction

Multi-view data generated from various view-points or multiple sensors are commonly seen in real-world applications. For example, the popular commercial depth sensor Kinect uses both visible light and near infrared sensors for depth estimation; autopilot uses both visual and radar sensors to produce real-time 3D information on the road; face analysis algorithms prefer face images from different views for high-fidelity reconstruction and recognition. However, such data with large *view* divergence would lead to an enormous challenge: data across various views have a large divergence preventing them from a fair comparison. Generally, different views tend to be treated as different domains from different distributions. Thus, there is an urgent need to mitigate the view divergence when facing specific problems by either fusing the knowledge across multiple views or adapting knowledge from some views to others. Since there are different terms regarding "multi-view" data analysis and its aliasing, we first give a formal definition and narrow down our research focus to differentiate it from other related works but in different lines.

Definition [Multi-view Data]: Assume we have a set of data $\mathcal{X} = \{X_1, X_2, \cdots, X_v\}$ from v views, e.g., face poses, camera views and types of features. In this paper, we are especially interested in two cases upon *data correspondence*: First, the samples across v views are correspondent (i.e., sample-wise relationship) in multi-view data, falling in the conventional **multi-view learning**; Second, the samples across different views have no data correspondence, falling in the **domain adaption** scenario, where discriminant knowledge are transferred.

First, multi-view learning aims to to merge the knowledge from different views to either uncover common knowledge, or employ the complementary knowledge in specific views to assist learning tasks, e.g., clustering [Zhao *et al.*, 2017; Tao *et al.*, 2017], outliers detection [Zhao *et al.*, 2018] and classification [Ding and Fu, 2014; 2017b; Kan *et al.*, 2016b; 2016a; Li *et al.*, 2017; Ding and Fu, 2016]. For example, in vision, multiple features extracted from the same object by various visual descriptors, e.g., LBP, SIFT and HOG are very discriminant in recognition tasks. Another example is multi-modal data captured, represented, and stored in varied formats, e.g., near-infrared & visible face, and image & text.

Second, domain adaptation attempts to transfer knowledge from labeled source domains to facilitate the learning burden in the target domains with sparsely or no labeled samples. For example, in surveillance, faces are captured by long wave infrared sensor in night-time, but recognition model is trained on regular face images collected under visible light. Conventional domain adaptation methods [Ding and Fu, 2017b; Ding *et al.*, 2015a; Shao *et al.*, 2014] consider seeking domain-invariant representation for the data or modifying classifiers to fight off the marginal or conditional distribution mismatch across source and target domains.

In this work, we provide a comprehensive review on robust multi-view data representation by jointly considering multi-view learning and domain adaptation as a unified learning framework. To the best of our knowledge, this is the first work to discuss both of them in such a unified perspective. Beyond their similarity, we further discuss their differences based on data organization and problem setting, as well as the

research goal. To sum up, we have our two-fold contributions as follows:

- First of all, we formulate multi-view learning and domain adaptation as a unified objective into two parts: multi-view alignment term and feature learning regularizer. This formulation would cover most multi-view representation learning algorithms in the fields of multi-view learning and domain adaptation.
- Secondly, based on the unified perspective, we further broaden our discussions of multi-view learning and domain adaptation, specifically for their different problem settings. Then, we lead a comprehensive review of our past research and other highly related work in this line.

2 A Unified Perspective

Due to the distribution divergence across different views, view-invariant feature learning is a widely-used and promising technique to address the multi-view challenges. Generally, multiple view-specific linear or non-linear mapping functions would be sought to transform the original multi-view data into a new common space by identifying dedicated alignment strategies with various loss functions. Specifically, we could formulate them into a common objective including two parts: (1) multi-view alignment term; (2) feature learning regularizer, namely:

$$\min_{f_1(\cdot),\cdots,f_v(\cdot)} \sum_{i=1,i<j}^{v} \mathcal{A}(f_i(X_i), f_j(X_j)) + \lambda \sum_{k=1}^{v} \mathcal{R}(f_k(X_k)),$$

where $f_i(\cdot)$ is a feature learning function for view i, either linear, non-linear mapping, or deep network.

The first common term $\mathcal{A}(\cdot)$ is a pairwise symmetric alignment function across multiple views to either fuse the knowledge among multiple views or transfer knowledge across different views. Due to different problem settings, multi-view learning and domain adaptation would explore various strategies to define the loss functions. While multi-view learning employs data correspondence (i.e., sample-wise relationship w/ or w/o labels) to seek common representation, domain adaptation employs domain- or class-wise relationship during the model learning for discriminant domain invariant feature.

The second common term $\mathcal{R}(\cdot)$ is the feature learning regularizer by incorporating either the labeled information or the intrinsic structure of the data, or both during the mapping learning. For a part of multi-view learning algorithms, they would merge feature learning regularizer into the alignment term. Generally, the formulation of the second term is very similar between multi-view learning and domain adaptation within our research concentration.

For clarity, Table 1 lists the frequently used notations.

3 Multi-View Learning

For multi-view learning, the goal is to fuse the knowledge from multiple views to facilitate common learning tasks, e.g., clustering and classification. The key challenge is exploring data correspondence across multiple views. The mappings among different views are able to couple view-specific knowledge while additional labels would help formulate supervised regularizers.

Notation	Description
$\|\cdot\|_{\mathrm{F}}$	Frobenius norm of a matrix
$\mathrm{rank}(\cdot)$	rank operator of a matrix
$\|\cdot\|_*$	Nuclear norm of a matrix
$\mathrm{tr}(\cdot)$	the trace of a matrix
$\|\cdot\|_1$	the l_1-norm of a matrix
$\|\cdot\|_{2,1}$	the $l_{2,1}$-norm of a matrix
P/W	View-invariant linear projection/rotation
P_i/W_i	the linear projection/rotation for view i
H	the new shared representation for all views
H_i	the new representation for the i-th view
L	a pre-defined matrix for all views
L_i	a pre-defined matrix for the i-th view
Z	the reconstruction matrix for all views
Z_i	the reconstruction matrix for the i-th view
S_w^i	within-class scatter matrix for the i-th view
S_w	within-class scatter matrix for all views
S_b^i	between-class scatter matrix for the i-th view
S_b	between-class scatter matrix for all views
$\alpha, \beta, \lambda, \eta$	balance parameters

Table 1: Notations and Descriptions.

3.1 Multi-view Clustering

The general setting of multi-view clustering is to group n data samples in v different views (e.g., v types of features, sensors, or modalities) by fusing the knowledge across different views to seek a consistent clustering result.

As known to all, canonical correlation analysis (CCA) is the most popular algorithm to achieve a common space for two views. Specifically, CCA attempted to obtain two projections, one for each view, to transform the data from two different views into a shared subspace, respectively, through maximizing the cross correlation across two views:

$$\mathcal{A}(\cdot) = -\mathrm{tr}(P_1^\top X_1 X_2^\top P_2)$$

with constraint as $\mathrm{tr}(P_i^\top X_i X_i^\top P_i) = \mathrm{I}_p$ $(i = 1, 2)$.

While pairwise constraint for multiple views works well, it becomes trivial when the number of views is large. Thus, given more views, a more efficient solution (Multi-view CCA [Chaudhuri *et al.*, 2009]) was proposed to seek a unified common space shared across all views:

$$\mathcal{A}(\cdot) = -\sum_{i,j=1,i\neq j}^{v} \mathrm{tr}(P_i^\top X_i X_j^\top P_j)$$

with constraint as $\mathrm{tr}(P_i^\top X_i X_i^\top P_i) = \mathrm{I}_p$ $(i = 1, \cdots, v)$.

Moreover, several recently proposed approaches for multi-view representation learning were based on deep neural networks (DNN), inspired by their success in typical unsupervised (single-view) feature learning settings. Andrew et al. proposed a DNN extension of CCA termed deep CCA (DCCA [Andrew *et al.*, 2013], where two DNNs are used to learn nonlinear features from each view and meanwhile the canonical correlation across two views is maximized:

$$\mathcal{A}(\cdot) = -\mathrm{tr}\Big(P_1^\top f_1(X_1) f_2(X_2)^\top P_2\Big)$$

with constraints as $\mathrm{tr}(P_i^\top(f_i(X_i)f_i(X_i)^\top)P_i) = \mathrm{I}_p$ $(i = 1, 2)$. $p_{1,k}^\top f_1(X_1)f_2(X_2)^\top p_{2,l} = 0$ for $k \neq l$, where $p_{i,k}$ is the k-th column of P_i.

Following the idea of deep learning, we explored semi-nonnegative matrix factorization to obtain the hierarchical semantics from multi-view data in a layer-wise manner [Zhao *et al.*, 2017]. To maximize the mutual information of each view, we couple new representations of all views to be the same in the final layer. Furthermore, graph regularizers are introduced to incorporate intrinsic geometric structure for the deep structures:

$$\begin{aligned}\mathcal{A}(\cdot) = &\sum_{i=1}^{v}\mathrm{tr}(HL_iH^\top),\\ \mathcal{R}(\cdot) = &\sum_{i=1}^{v}(\alpha_i)^\gamma\|X_i - P_i^1P_i^2\cdots P_i^mH\|_{\mathrm{F}}^2\\ = &\sum_{i=1}^{v}(\alpha_i)^\gamma\|f_i(X_i) - H\|_{\mathrm{F}}^2\end{aligned}$$

with constraints $H \geq 0, \sum_{i=1}^{v}\alpha_i = 1, \alpha_i \geq 0$. And γ is adopted to balance the weights distribution.

Discussions: As no labels are available, the common representation by either deep or shallow structure is learned based on underlying distributions or the descriptors of data. The statistics of data together with deep representation learning plays a critical role. However, how to extend such as numerical methods to large-scale dataset will be an open question.

3.2 Multi-view Classification

The general setting of multi-view classification is that it needs to build a model with given v views of training data. In the test stage, we would have two different scenarios. First, one view will be used to recognize other views with the learned model. In this case, the label information across training and test data is different[1]; Second, specifically for multi-features based learning, is that v-view training data is used to seek a model by fusing the cross-view knowledge, which is also used as gallery data to recognize v-view probe data[2].

Traditionally, to seek a discriminant shared space for multiple views, the label information is usually incorporated. Along this line, Sharma et al. developed a generalized multi-view analysis (GMA) framework [Sharma *et al.*, 2012], where the supervised information was involved as:

$$\mathcal{A}(\cdot) = \sum_{i=1}^{v}\mu_i\mathrm{tr}(P_i^\top S_b^iP_i) + \sum_{i<j}^{v}\lambda_{ij}\mathrm{tr}(P_i^\top X_iX_j^\top P_j)$$

with constraints $\sum_{i=1}^{v}\gamma_i\mathrm{tr}(P_i^\top S_w^iP_i) = \mathrm{I}_p$.

Moreover, Cai et al. adopted partial least squares (PLS) regression to classify faces with variations in pose, resolution and image domains [Cai *et al.*, 2013]. To better align multiple view-specific transformations, they employed a smooth constraint as well:

$$\begin{aligned}\mathcal{A}(\cdot) &= \sum_{i=1}^{v}\|HL_i - P_iX_i\|_{\mathrm{F}}^2 + \beta\sum_{i=1}^{v-1}\|P_i - P_{i+1}\|_{\mathrm{F}}^2,\\ \mathcal{R}(\cdot) &= \mathrm{tr}(HLH^\top),\end{aligned}$$

where they only explored the constraints on the poses in their neighborhoods.

Although GMA and PLS are able to learn a discriminant common subspace, they only took into consideration the within-view knowledge, but ignored the between-view information. To this end, Kan et al. proposed a multi-view discriminant analysis approach [Kan *et al.*, 2016b] that can achieve a single discriminant shared space for all views by simultaneously learning v view-specific projections:

$$\mathcal{A}(\cdot) = \frac{\mathrm{tr}(P_xS_wP_x^\top)}{\mathrm{tr}(P_xS_bP_x^\top)}, \tag{1}$$

where $P_x = [P_1^\top X_1, \cdots, P_v^\top X_v]$.

By exploring deep neural networks, Kan et al. further proposed a multi-view deep network, which aims to eliminate the complex view divergence for favorable multi-view learning by seeking deep view-invariant and discriminant features [Kan *et al.*, 2016a]. Specifically, the MvDN architecture includes two sub-networks, one is view-specific sub-network $f_i(\cdot)$ to reduce view-specific variations and the other is common sub-network $g_c(\cdot)$ to seek shared representation across all views. Finally, they adopted the Fisher loss, i.e. the Rayleigh quotient objective, to guide the whole architecture learning:

$$\mathcal{A}(\cdot) = \frac{\mathrm{tr}(F_xS_wF_x^\top)}{\mathrm{tr}(F_xS_bF_x^\top)},$$

where $F_x = [g_c(f_1(X_1)), \cdots, g_c(f_v(X_v))]$.

Previously mentioned algorithms were designed to solve general multi-view classification problems. Thus, they mainly consider label information during the multi-view alignment. For applications with domain knowledge, e.g., kinship verification and action recognition, specific loss functions may be required. For kinship verification, we developed a Coupled Marginalized Auto-Encoders [Wang *et al.*, 2016], where two marginalized denoising auto-encoders were designed for source and target views. To better align two denoising auto-encoders, a feature mapping was incorporated to adapt knowledge across the intermediate and the target view. Furthermore, the maximum margin criterion is imposed on the top layer to learn more discriminative representations across those domains as:

$$\begin{aligned}\mathcal{A}(\cdot) = &\|P_1^\top X_1 - WP_2^\top X_2\|_{\mathrm{F}}^2 +\\ &\alpha\sum_{k,l=1}^{n}y_{kl}\|P_1P_1^\top X_1^k - P_2WP_2^\top X_2^l\|_2^2,\\ \mathcal{R}(\cdot) = &\sum_{i=1}^{2}\|\bar{X}_i - P_iP_i^\top\tilde{X}_i\|_{\mathrm{F}}^2,\end{aligned}$$

where $y_{kl} \in \{-1, 1\}$ indicates the relationship between the k-th sample of X_1 and the l-th sample of X_2, either positive (1) or negative (-1).

[1] Zero-shot learning [Ding *et al.*, 2017] can also be modeled as a special case of multi-view learning, which involves two views, i.e., visual features and semantic features.

[2] In this second case, gallery data and probe data have the same label information. In reality, we may confront such a challenge that we have no prior knowledge for the view information of test data, especially for cross-pose and cross-modality image recognition [Ding and Fu, 2014; 2017a].

We also explored the cross-view action recognition problem in this line by seeking view-specific and view-shared networks through novel deep models [Kong *et al.*, 2017]. Specifically, view-specific networks target at capturing unique patterns within each view, and view-shared network aim to encode common dynamics across different views. Furthermore, we explored the incoherence across the two types of networks, which is encouraged to remove information redundancy and uncover more discriminant knowledge:

$$\mathcal{A}(\cdot) = \sum_{i=1}^{v} \left(\|W^\top W_i\|_{\mathrm{F}}^2 + \beta \mathrm{tr}(W_i X_i L X_i^\top W_i^\top) \right),$$
$$\mathcal{R}(\cdot) = \|W\tilde{X} - XL\|_{\mathrm{F}}^2 + \sum_{i=1}^{v} \alpha \|W_i \tilde{X}_i - X_i\|_{\mathrm{F}}^2.$$

Compared with traditional neural networks [Kan *et al.*, 2016a] and auto-encoder [Andrew *et al.*, 2013], marginalized denoising auto-encoder adopted by [Wang *et al.*, 2016; Kong *et al.*, 2017; Ding *et al.*, 2015a] is much faster and also achieves comparable results in many applications. Thus, it is widely used in various large-scale problems recently.

Discussions: The success of these methods partially lies in the discriminant features for either face or video analysis as the methods above primarily concentrate on high-level feature modeling. Thus, Fisher criterion or fast running marginalized auto-encoder could be applied. Nonetheless, in face of unbalanced data across classes, or poor features, these modeling would fail. Incorporating robustness and end-to-end discriminant feature learning will be our future research.

In our previous work, we also explored some challenging problems in multi-view classification, e.g., zero-shot learning and view-unknown learning problem.

Zero-shot Learning

There is a special case of multi-view learning, i.e., zero-shot learning (ZSL), which is inspired by the learning mechanism of human brain. The goal is to classify new categories which are unobserved during the training process. For instance, one is able to predict a new species of animal after being informed what it looks like and how it is different from or similar to other known animals. Generally, there will be two views in ZSL, i.e., visual features and semantic features that are highly coupled. Different from conventional multi-view classification, the key of ZSL is to find the mapping across two views and generalize well to unseen test data.

To this end, we developed an effective Low-rank Embedded ensemble Semantic Dictionary learning to solve zero-shot learning [Ding *et al.*, 2017], with the main assumption as the latent semantic dictionary learned from the seen categories should contain majority information for the unseen categories. Furthermore, we exploited multiple transferable dictionaries through ensemble strategy to have a better chance to recover the latent semantic space for unseen data:

$$\mathcal{A}(\cdot) = \sum_{k=1}^{K} \|W X_1^k - D X_2^k Z_k\|_{\mathrm{F}}^2,$$
$$\mathcal{R}(\cdot) = \mathrm{rank}(W) \rightarrow \sum_{i=r+1}^{d} \sigma_i^2(W),$$

where $X_{1/2}^k$ is a random selection of $X_{1/2}$ and $\sigma_i(W)$ is the i-th smallest singular value.

When Probe View is Unknown

Traditional multi-view learning approaches targeted at seeking multiple view-specific projections either linear or non-linear, since they assumed the view information of training and test data were already accessible in advance. Actually, we always confront the situations that we have no prior for the test data's view information, and therefore, multiple view-specific projections cannot be used to learn its specific feature representations. To this end, we proposed to seek a view-invariant projection to fight off this challenge in multi-view data learning scenarios [Ding and Fu, 2014; 2017a]:

$$\mathcal{A}(\cdot) = \sum_{i=1}^{v} \left(\|Z_i\|_* + \alpha \|P_i - P\|_1 \right) + \beta \mathrm{tr}(\mathcal{S}_w - \eta \mathcal{S}_b)$$

with constraints as $P_i^\top X_i = P^\top D Z_i$, $P^\top P = \mathrm{I}_p$ and $\mathcal{S}_{w/b}$ is the within-class/between-class scatter matrix defined on $P^\top D[Z_1, \cdots, Z_v]$. In our model, we attempt to address the semantic gap across multiple views by learning a shared transformation from multiple view-specific ones. To achieve this, low-rank reconstruction is explored to bridge the view-specific features and the view-invariant ones transformed with the collective low-rank subspace. Furthermore, we adopted a supervised cross-view regularizer to align the intra-class data across multiple views.

Discussions: While these methods are promising in the prediction tasks with semantic gap, it usually showcases between similar visual concepts, e.g., different objects, categories of animals, or different views of same subject. A more general learning paradigm for intelligent recognition targeting at large semantic gaps will be the future direction.

4 Domain Adaption

The goal of domain adaptation is to transfer knowledge from well-labeled sources to unlabeled targets, which accounts for the more general settings that some source views are labeled while target views are unlabeled. The general setting of domain adaptation is that we build a model on both labeled source data X_s and unlabeled target data X_t. Then we use the model to predict the unlabeled target data, either the same data in the training stage or different data. Thus, we have corresponding transductive domain adaptation and inductive domain adaption.

To make domain adaption more general, we define $v_s (v_s \geq 1)$ source data $X_s = \{X_1|Y_1, \cdots, X_{v_s}|Y_{v_s}\}$ and $v_t (v_t \geq 1)$ unlabeled target data $X_t = \{X_{v_s+1}|?, \cdots, X_v|?\}$ $(v = v_s + v_t)$. The task is to train a model on labeled source and unlabeled target to predict the label of target data. A feasible and practical way is joint feature learning and domain alignment to seek more effective domain-invariant space. For simplicity, we also denote $\bar{X}$ as the m copy of $X = [X_s, X_t]$, and $\tilde{X}$ as the corruption of $\bar{X}$.

Currently, reconstruction error [Ding *et al.*, 2016; Li *et al.*, 2016; Shao *et al.*, 2014] and Maximum Mean Discrepancy (MMD) [Ding and Fu, 2017b] are two promising techniques as the distance measure to compare different distributions of source and target domains.

4.1 Transfer Subspace Learning

Subspace learning is the most popular feature learning strategy. Along this line, we proposed a Transfer Subspace Learning [Shao *et al.*, 2014] by transforming both source and target data into a domain-invariant subspace, in which each target sample will be reconstructed by several source samples from a neighborhood. Furthermore, low-rank constraint was adopted to guide the reconstruction, and therefore, this knowledge transfer scheme can preserve the intrinsic structures of source and target domains. Through an iterative optimization way, good alignment across two domains tends to be guaranteed when the target samples are only remonstrated by several relevant samples of the source domain in the latent space, ideally the same-class data. Then the discriminability in the source domain will be naturally passed on to the target domain. Specifically, the two parts in [Shao *et al.*, 2014] are defined as follows with constraint $P^\top P = \mathrm{I}_p$:

$$\begin{aligned}\mathcal{A}(\cdot) &= \|P^\top X_s - P^\top X_t Z\|_{\mathrm{F}}^2 + \beta \mathrm{rank}(Z),\\ \mathcal{R}(\cdot) &= \mathrm{tr}(P^\top X_s L X_s^\top P).\end{aligned}$$

Later on, Li et al. developed a domain adaption framework, which smoothly merges feature selection as well as structure preservation into a unified model [Li *et al.*, 2016]. Specifically, the two parts in [Li *et al.*, 2016] are defined as follows with constraint $P^\top P = \mathrm{I}_p$:

$$\begin{aligned}\mathcal{A}(\cdot) &= \|P^\top X - P^\top X_s Z\|_{\mathrm{F}}^2 + \beta \|Z\|_{\mathrm{F}}^2,\\ \mathcal{R}(\cdot) &= \mathrm{tr}(P^\top X L X^\top P) + \alpha \|P\|_{2,1},\end{aligned}$$

which is different from our low-rank transfer learning [Shao *et al.*, 2014; Ding *et al.*, 2015b] as it explored Frobenius norm to replace rank constraint and further exploited a feature selection regularizer on subspace projection. Thus, it can speed up the optimization without rank constraint, and deploy group sparsity to seek a more effective subspace. Similarly, we also explored Frobenius norm to guide the reconstruction for knowledge transfer to segment human motion in an unsupervised fashion [Wang *et al.*, 2018].

Furthermore, Tsai et al. particularly addressed the practical and challenging scenario of imbalanced cross-domain data [Tsai *et al.*, 2016]. That is, the label numbers across domains are not assumed to be the same. To solve the above task of imbalanced domain adaptation, they proposed a novel algorithm of domain-constraint transfer coding with constraint $P^\top X L_1 X^\top P = \mathrm{I}_p$:

$$\mathcal{A}(\cdot) = \|P^\top X_t - P^\top X_s Z\|_{\mathrm{F}}^2 + \beta \|L \odot Z\|_{\mathrm{F}}^2,$$

which is able to exploit latent sub-domains within and across data domains, and learns a common feature space for joint adaptation and classification purposes. $\odot$ is element-wise matrix multiplication.

The downside of previous works is they all ignored conditional distribution, which motivates us to jointly consider marginal and conditional distributions by a robust knowledge transfer metric [Ding and Fu, 2017b]. Specifically, we exploit knowledge transfer to mitigate the domain shift in two directions, i.e., sample space and feature space (Note that $\mathcal{M} = PP^\top$ is a semi-positive definite matrix):

$$\mathcal{A}(\cdot) = \mathrm{tr}(\mathcal{S}\mathcal{M}), \quad \mathcal{R}(\cdot) = \|\bar{X} - \mathcal{M}\tilde{X}\|_{\mathrm{F}}^2 + \alpha \mathrm{rank}(\mathcal{M}),$$

where $\mathcal{S}$ is domain/class-wise mean difference matrix. Similarly, we also boost the our previous reconstruction-based domain adaptation models through conditional distribution matching and adopt a pre-defined or iterative-updated structure matrix to guide the reconstruction coefficients learning in [Ding *et al.*, 2015a; 2016; Ding and Fu, 2018].

Missing Modality Transfer Learning

An interesting problem in multi-modality learning is lack of target data, i.e., one specific modality, in training stage. Fortunately, we can borrow the knowledge from a complete multi-modality dataset and model it as domain adaptation in two directions, i.e., cross-modality and cross-dataset. To that end, we designed a novel framework by exploring two-directional transfer [Ding *et al.*, 2014; 2015b], each of which is defined as follows with constraint $P^\top P = \mathrm{I}_p$:

$$\begin{aligned}\mathcal{A}(\cdot) &= \|Y_s - P^\top X_t Z + W Y_s\|_{\mathrm{F}}^2 + \alpha(\|Z\|_* + \|W\|_*),\\ \mathcal{R}(\cdot) &= \|P\|_{2,1} + \beta \mathrm{tr}(P^\top L P),\end{aligned}$$

where a latent factor W is generated to seek the underlying structure of the missing modality from the observed modalities. Thus, we iteratively consider two-directional transfer, which allows the knowledge transfer across both modalities and databases to mitigate the missing modality. Note this may be conceptually similar to zero-shot learning, or domain generalization while we are able to refer to an auxiliary dataset in this case.

Incomplete Multi-Source Transfer Learning

In recent problems, multiple sources may account for knowledge adaption, however, each one may not contain complete categories information compared to the target domain. Simply merging multiple sources as a whole would result in inferior output because of the large discrepancy within multiple sources. Thus, we aim to explore better knowledge transfer from incomplete multiple sources to boost the learning task for target domain [Ding *et al.*, 2016]. Finally, we developed an incomplete multi-source transfer subspace learning algorithm from through two directions, one is cross-domain knowledge transfer from each source to target domain, where we deploy a latent low-rank transfer scheme [Ding *et al.*, 2015b; 2014] to implicitly recover the missing categories in each source; and the other is cross-source knowledge transfer to joint multi-source information effectively, where we design an unsupervised graph term to couple multiple sources in order to compensate for incomplete categories from one source to another. With the orthogonal constraint $P^\top P = \mathrm{I}_p$, we have the two terms as:

$$\begin{aligned}\mathcal{A}(\cdot) = \sum_{i=1}^{v_s} \big(&\|Z_i\|_* + \|W_i\|_* + \alpha \|Z_i - L_i\|_{\mathrm{F}}^2\\ &+ \beta \|P^\top X_t - Y_{s,i} Z_i - W_i P^\top X_t\|_{\mathrm{F}}^2\big)\\ \mathcal{R}(\cdot) = \sum_{i=1}^{v_s} &\mathrm{tr}(Y_{s,i} Z_i L Z_i^\top Y_{s,i}^\top) + \mathrm{tr}(P^\top X_t L X_t^\top P).\end{aligned}$$

Discussions: As widely discussed along with subspace learning, kernelization and tensorlization are better counterparts for non-parametric and multi-linear modeling. While deep learning plays key roles in methods above, they will not work well given limited data. Thus, there is a need to develop kernel or tensor methods in this line. In addition, most of transfer subspace learning in this vein requires large matrices products and eigen-decomposition, which imposes additional computing load with large dataset. Thus, algorithm accelerating would benefit the deployment in the real-world applications, especially when dealing with large-scale data.

4.2 Deep Domain Adaption

Recently, we explored a stacked deep low-rank coding framework [Ding *et al.*, 2015a] for knowledge transfer. Specifically, for each layer, we obtained discriminative low-rank coding with the guidance of marginalized denoising strategy and an iterative structured term. Hence, both marginal and conditional differences across two domains can be well mitigated. The two parts for each layer are defined as follows:

$$\mathcal{A}(\cdot) = \|WX - WX_sZ\|_{\mathrm{F}}^2 + \alpha\|Z\|_* + \beta\|Z - L\|_{\mathrm{F}}^2,$$
$$\mathcal{R}(\cdot) = \mathrm{tr}\big[(\bar{X} - W\tilde{X})^\top(\bar{X} - W\tilde{X})\big],$$

where we could further achieve next-layer coding with the learned previous-layer coding in a layer-wise fashion.

Most recently, deep domain adaptation [Long *et al.*, 2015; Rozantsev *et al.*, 2018] targeted at improving the feature adaptation ability in the top layers of DNNs by explicitly mitigating the domain shift. Therefore, they are able to achieve feed-forward architectures, which are applicable to the target domain without being harmed by the domain mismatch. Specifically, they adopt the deep architectures, e.g., AlexNet, GoogLeNet, ResNet, with domain alignment constraint (e.g., MMD or CORAL [Sun *et al.*, 2016]) at the top layers. Moreover, cross-entropy loss on the labeled source data is often adopted as the feature learning regularizer.

Domain Generalization

Existing domain adaptation algorithms all assumes that target data are still available for training although they are unlabeled. However, it would always happen in reality that the target data are totally inaccessible in advance. This is extremely challenging since we have no prior knowledge of the target domain. To fight off this issue, we developed a deep domain generalization algorithm by seeking consistent knowledge from multiple available source domains, where we explored a structured low-rank reconstruction to guide the knowledge transfer from each source to the unseen target domain as follows:

$$\mathcal{A}(\cdot) = \sum\nolimits_{i=1}^{v} \|H_i - HZ_i\|_{\mathrm{F}}^2 + \|Z\|_* + \alpha\|Z - L\|_{\mathrm{F}}^2,$$

where we designed multiple domain-specific DNNs to learn the rich knowledge within multiple source domains, and simultaneously a domain-shared DNNs to capture the common information across multiple sources. In this way, such a domain-shared DNN is still valid to unseen target domain.

Discussion: There is always a debate in this line whether an end-to-end deep learning paradigm is needed. While some research reports better results using end-to-end training, the performance is evaluated on small-scale datasets. Domain adaptation usually offers small amount of target data, either labeled or not, and thus, fine-tuning across large semantic gaps may not work well. Therefore, deep features with conventional domain adaptation methods is not a bad choice.

5 Conclusions and Future Work

In this paper, we presented a comprehensive survey on robust data representation for multi-view learning and domain adaptation problems. We identified the shared and distinct terms across multi-view learning and domain adaptation, and lead a detailed discussion including our recently proposed algorithms for multi-view clustering, multi-view classification and domain adaptation in general, and zero-shot learning, view-unknown learning, missing modality learning, and incomplete multi-source learning in particular. This would benefit the AI community in both industry and academia from literature review to future directions. Despite the recent advances, in future research, we will focus on the following factors: imbalanced, incomplete, and large-scale datasets, as identified in our previous discussions:

First, large-scale multi-view image retrieval needs many image pairs across views to learn correspondence. But both the probe and reference images are not under control in terms of both quality and quantity. For example, in forensic face recognition, we have a single sketch face as reference to retrieve RGB faces from surveillance cameras, and enrolled face images from police department. The sketch needs to be converted to common feature first and then compared against RGB faces. The single sketch and many other RGB faces from different persons, with varied quality and numbers pose an extreme unbalanced learning.

Second, how to adapt the knowledge from existing large-scale public datasets to new domains or problems where training samples are few? This is extremely critical for problems that need knowledge extrapolation. This is essentially a "compound" of few-shot learning and domain adaptation. We still take face recognition as an example, where we intend to extend the well-trained face recognition algorithms for day time to night light under poor illuminations. We may only given few images per person in night time, which accounts for the extremely incomplete multi-view data.

Finally, generalizing the discussed methods to large-scale datasets in the wild is the ultimate goal as most of them require intensive computing in numerical optimization, e.g., $O(n^3)$ where n is number of the samples. We may refer to existing efficient solutions for eigen-decomposition that shrink it down to $O(n^2)$ under mild condition, or other heuristics including "divide and conquer". We will dedicate to the toolbox development and benchmarks for robust multi-view representation learning in this line.

Acknowledgements

This research is supported in part by the NSF IIS award 1651902, ONR Young Investigator Award N00014-14-1-0484, and U.S. Army Research Office Award W911NF-17-1-0367.

References

[Andrew *et al.*, 2013] Galen Andrew, Raman Arora, Jeff Bilmes, and Karen Livescu. Deep canonical correlation analysis. In *ICML*, pages 1247–1255, 2013.

[Cai *et al.*, 2013] Xinyuan Cai, Chunheng Wang, Baihua Xiao, Xue Chen, and Ji Zhou. Regularized latent least square regression for cross pose face recognition. In *IJCAI*, pages 1247–1253. AAAI Press, 2013.

[Chaudhuri *et al.*, 2009] Kamalika Chaudhuri, Sham M Kakade, Karen Livescu, and Karthik Sridharan. Multi-view clustering via canonical correlation analysis. In *ICML*, pages 129–136. ACM, 2009.

[Ding and Fu, 2014] Zhengming Ding and Yun Fu. Low-rank common subspace for multi-view learning. In *ICDM*, pages 110–119. IEEE, 2014.

[Ding and Fu, 2016] Zhengming Ding and Yun Fu. Robust multi-view subspace learning through dual low-rank decompositions. In *AAAI*, pages 1181–1187. AAAI Press, 2016.

[Ding and Fu, 2017a] Zhengming Ding and Yun Fu. Robust multiview data analysis through collective low-rank subspace. *IEEE Transactions on Neural Networks and Learning Systems*, 2017.

[Ding and Fu, 2017b] Zhengming Ding and Yun Fu. Robust transfer metric learning for image classification. *IEEE Transactions on Image Processing*, 26(2):660–670, 2017.

[Ding and Fu, 2018] Zhengming Ding and Yun Fu. Deep domain generalization with structured low-rank constraint. *IEEE Transactions on Image Processing*, 27(1):304–313, 2018.

[Ding *et al.*, 2014] Zhengming Ding, Ming Shao, and Yun Fu. Latent low-rank transfer subspace learning for missing modality recognition. In *AAAI*, pages 1192–1198. AAAI Press, 2014.

[Ding *et al.*, 2015a] Zhengming Ding, Ming Shao, and Yun Fu. Deep low-rank coding for transfer learning. In *IJCAI*, pages 3453–3459. AAAI Press, 2015.

[Ding *et al.*, 2015b] Zhengming Ding, Ming Shao, and Yun Fu. Missing modality transfer learning via latent low-rank constraint. *IEEE Transactions on Image Processing*, 24(11):4322–4334, 2015.

[Ding *et al.*, 2016] Zhengming Ding, Ming Shao, and Yun Fu. Incomplete multisource transfer learning. *IEEE Transactions on Neural Networks and Learning Systems*, 2016.

[Ding *et al.*, 2017] Zhengming Ding, Ming Shao, and Yun Fu. Low-rank embedded ensemble semantic dictionary for zero-shot learning. In *CVPR*, pages 2050–2058, 2017.

[Kan *et al.*, 2016a] Meina Kan, Shiguang Shan, and Xilin Chen. Multi-view deep network for cross-view classification. In *CVPR*, pages 4847–4855, 2016.

[Kan *et al.*, 2016b] Meina Kan, Shiguang Shan, Haihong Zhang, Shihong Lao, and Xilin Chen. Multi-view discriminant analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 38(1):188–194, 2016.

[Kong *et al.*, 2017] Yu Kong, Zhengming Ding, Jun Li, and Yun Fu. Deeply learned view-invariant features for cross-view action recognition. *IEEE Transactions on Image Processing*, 26(6):3028–3037, 2017.

[Li *et al.*, 2016] Jingjing Li, Jidong Zhao, and Ke Lu. Joint feature selection and structure preservation for domain adaptation. In *IJCAI*, pages 1697–1703, 2016.

[Li *et al.*, 2017] Jingjing Li, Yue Wu, Jidong Zhao, and Ke Lu. Low-rank discriminant embedding for multiview learning. *IEEE Transactions on Cybernetics*, 47(11):3516–3529, 2017.

[Long *et al.*, 2015] Mingsheng Long, Yue Cao, Jianmin Wang, and Michael Jordan. Learning transferable features with deep adaptation networks. In *ICML*, pages 97–105, 2015.

[Rozantsev *et al.*, 2018] Artem Rozantsev, Mathieu Salzmann, and Pascal Fua. Beyond sharing weights for deep domain adaptation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2018.

[Shao *et al.*, 2014] Ming Shao, Dmitry Kit, and Yun Fu. Generalized transfer subspace learning through low-rank constraint. *International Journal of Computer Vision*, 109(1-2):74–93, 2014.

[Sharma *et al.*, 2012] Abhishek Sharma, Abhishek Kumar, Hal Daume, and David W Jacobs. Generalized multiview analysis: A discriminative latent space. In *CVPR*, pages 2160–2167. IEEE, 2012.

[Sun *et al.*, 2016] Baochen Sun, Jiashi Feng, and Kate Saenko. Return of frustratingly easy domain adaptation. In *AAAI*, pages 2058–2065. AAAI Press, 2016.

[Tao *et al.*, 2017] Zhiqiang Tao, Hongfu Liu, Sheng Li, Zhengming Ding, and Yun Fu. From ensemble clustering to multi-view clustering. In *IJCAI*, pages 2843–2849. AAAI Press, 2017.

[Tsai *et al.*, 2016] Yao-Hung Hubert Tsai, Cheng-An Hou, Wei-Yu Chen, Yi-Ren Yeh, and Yu-Chiang Frank Wang. Domain-constraint transfer coding for imbalanced unsupervised domain adaptation. In *AAAI*, pages 3597–3603. AAAI Press, 2016.

[Wang *et al.*, 2016] Shuyang Wang, Zhengming Ding, and Yun Fu. Coupled marginalized auto-encoders for cross-domain multi-view learning. In *IJCAI*, pages 2125–2131. AAAI Press, 2016.

[Wang *et al.*, 2018] Lichen Wang, Zhengming Ding, and Yun Fu. Learning transferable subspace for human motion segmentation. In *AAAI*. AAAI Press, 2018.

[Zhao *et al.*, 2017] Handong Zhao, Zhengming Ding, and Yun Fu. Multi-view clustering via deep matrix factorization. In *AAAI*, pages 2921–2927, 2017.

[Zhao *et al.*, 2018] Handong Zhao, Hongfu Liu, Zhengming Ding, and Yun Fu. Consensus regularized multi-view outlier detection. *IEEE Transactions on Image Processing*, 27(1):236–248, 2018.