Can Visual Recognition Benefit from Auxiliary Information in Training?

Qilin Zhang[1], Gang Hua[1], Wei Liu[2], Zicheng Liu[3], Zhengyou Zhang[3]

[1]Stevens Institute of Technolog, Hoboken, NJ
[2]IBM Thomas J. Watson Research Center, Yorktown Heights, NY
[3]Microsoft Research, Redmond, WA

Abstract. We examine an under-explored visual recognition problem, where we have a main view along with an auxiliary view of visual information present in the training data, but merely the main view is available in the test data. To effectively leverage the auxiliary view to train a stronger classifier, we propose a collaborative auxiliary learning framework based on a new discriminative canonical correlation analysis. This framework reveals a common semantic space shared across both views through enforcing a series of nonlinear projections. Such projections automatically embed the discriminative cues hidden in both views into the common space, and better visual recognition is thus achieved on the test data that stems from only the main view. The efficacy of our proposed auxiliary learning approach is demonstrated through three challenging visual recognition tasks with different kinds of auxiliary information.

1 Introduction

We explore a new visual recognition problem dealing with visual data of two views, where a main view along with an auxiliary view is present in the training data. This particular vision problem attracts our attention, due to the recent popularity of multi-sensory and multiple spectrum imaging, such as depth-sensing and infrared (IR) sensing cameras, and hence the accumulation of labeled multi-view visual data.

However, a “missing-of-auxiliary view” problem frequently occurs in the test phase. This phenomenon could be incurred by sensor malfunction caused by, for example, an adversarial sensing environment or insufficient bandwidth allowing the transmission of only the main view data. In addition, the “missing view” problem could also arise when processing a backlog of historical data without an auxiliary sensing channel. Then a question naturally emerges: can visual recognition on the main view benefit from such auxiliary information that only exists in the training data?

Unlike conventional settings where training and testing data follow similar, if not identical, probability distributions [1], this problem requires techniques which can incorporate the beneficial information from the auxiliary view into the training of a classification model that works only with the main view. We shall emphasize that the problem studied in this paper is different from the domain adaptation and transfer learning problems [2–4] in computer vision. The goal of most domain adaptation/transfer learning problems is to leverage existing abundant labeled data in one domain to facilitate learning a better model in the target domain with scarce labeled data, if at all. Essentially, the knowledge is transferred from one data domain to a related but statistically different one. In contrast, in our problem the data domain in the test phase (*i.e.*, the main view) is a proper subset of the training data domain that contains the auxiliary view other than the main view.

In this sense, the problem we are addressing is more closely related to the multi-view and multi-task learning problems. The previous work in multi-view learning [5–7] has demonstrated that improved performance can be achieved by synergistically modeling all views. As a matter of fact, the existing multi-view recognition methods [5, 8, 9] emphasize heavily on properly combining per-view information.

We adopt a verification-by-construction approach by showing that there exists at least one method that consistently presents higher recognition accuracy on several multi-view visual datasets under this problem setting. In particular, we propose to seek a common semantic space to capture the relevance between the main and auxiliary views. This is achieved by a new discriminative canonical correlation analysis (DCCA) inspired by [10]. The new DCCA algorithm not only takes supervised label information into consideration, but also concurrently optimizes multiple nonlinear projections with a guaranteed convergence. Our DCCA algorithm is parallel to and even exceeds most previous CCA algorithms which did not explore label information and pursued multiple projections one by one.

With a desirable common semantic space, the auxiliary view information in the training data is carried into a classifier defined in the common space. Subsequent tests are conducted by projecting the only available main view information of the test data onto the common space and then applying the classifier obtained in the training phase.

The primary contributions of this paper are: (1) we focus on an under-explored visual recognition problem, *i.e.*, the “missing-view-in-test-data” problem with real-world multisensory visual datasets; (2) we propose a new discriminative canonical correlation algorithm together with a rigorous convergence proof, and its efficacy is validated on three benchmarks.

The rest of the paper is organized as follows. Section 2 briefly summarizes the related work. Section 3 formally defines the “missing-view-in-test-data” problem. Section 4 formulates the proposed DCCA algorithm. Section 5 gives a solution to the DCCA algorithm. Section 6 presents the experiments. Section 7 concludes the paper.

2 Related Work

Here we summarize the related work in domain adaptation, transfer learning, and multi-view learning, and also highlight their differences from the "missing-view-in-test-data" problem we investigate in the paper.

Recently, metric learning has been successfully applied to domain adaptation/transfer learning problems [1][3]. These methods attempt to transfer discriminative information from a source domain to a related but statistically d-ifferent target domain. Metric learning based domain transfer methods can be applied to various problems such as machine translation [5][2], multimedia information retrieval [11], and visual recognition [3][4].

It is noted that these methods assume abundant labeled training samples present in the source domain. While in the target domain, there are a limited number of labeled training samples. Therefore, it is difficult to directly train an effective classifier using the scarce labeled data in the target domain. Then the recent approaches such as [3][4] exploit the corresponding relationship between the two domains to build a regularized cross-domain transform via techniques such as metric learning and kernel learning [12] to fulfill knowledge transferring. However, these domain transfer problems are different from the "missing-view-in-test-data" problem we are tackling in the paper. In our problem, there exists a bijection between every corresponding pair of the (main,auxiliary)-view observations in the training set, due to the intrinsic semantic consistency between the two views.

The multi-view/multi-task learning, *e.g.*, [5][8][9], endeavors to learn a principled fusion which combines information from two or more related but statistically different views/tasks to achieve certain goals. It was demonstrated in [5] that the learning performance does benefit from explicitly leveraging the underlying semantic consistency between two views or among multiple views, which also motivates us to leverage this cross-view semantic consistency appearing in our problem. In our experiments shown in Section 6, a simple modification of the SVM2K algorithm [5] is implemented and treated as a competing baseline.

Another line of related work which can benefit from the multi-view data may be the co-training framework [13] proposed by Blum and Mitchell. Nevertheless, it falls into semi-supervised learning, and cannot deal with missing views. In [14], both the missing view and the domain transfer problem are considered, however, the objective function ignores the discriminative label information. A different RGBD-RGB data based object detection problem is addressed in [15] where explicit 3D geometry information is modeled. Tommasi *et al.* [16] focus on the dataset bias problem with a similar latent space model. In [17], a different missing feature problem is addressed, where the features are missing randomly in all dimensions, instead of the systematic absence of views in the multi-view settings considered in this paper. In the deep learning community, multi-modal deep learning systems such as [18] also show the robustness against missing views. In another related field, Chen *et al.* [19] suggest boosting-based learning with side information. Unlike our proposed latent space model, this method is not straightforward in terms of handling multiple sets of side information. In the area of human perception and psychology, Shams *et al.* [20] show that humans recognize better in unisensory tests with previous multisensory experiences than with the unisensory counterparts. Through establishing our proposed auxiliary learning framework, we show that this benefit also exists in the context of visual recognition. We also note that there exists a discriminative CCA [21] proposed for image set recognition where each sample consists of many observations/views. However, this CCA method exploits the correlations between sets of observations, which require abundant views for the robust estimation of sample covariance matrices. In contrast, our missing-view-in-test-data problem involves many observations but only a few views and even a single view in the test phase, so we are able to more easily exploit the correlations between views.

3 A Latent Space Model: Addressing Missing-View-in-Test-Data

Suppose that m view observations $\mathbf{z}(i)\overset{\text{def}}{=}(\mathbf{x}_1(i), \mathbf{x}_2(i))$ $(i = 1, 2, \cdots, m)$ are generated. Let $\mathbf{x}_1(i) \in \mathcal{X}_1$ and $\mathbf{x}_2(i) \in \mathcal{X}_2$ denote observations from the main view and auxiliary view, respectively. We assume $\mathcal{X}_1 \subset \mathbb{R}^{n_1}$ and $\mathcal{X}_2 \subset \mathbb{R}^{n_2}$, and define $\mathcal{Z}\overset{\text{def}}{=}\mathcal{X}_1 \times \mathcal{X}_2$, so $\mathbf{z}(i) \in \mathcal{Z}$. Within the training dataset, there comes with a label $l(i) \in \mathcal{L}$ for each observation $\mathbf{z}(i)$.

Unlike a conventional visual object recognition problem whose classifier $f_{\boldsymbol{\theta}} : \mathcal{Z} \to \mathcal{L}$ is obtained by the identification of the parameters $\boldsymbol{\theta} = \boldsymbol{\theta}(\mathbf{z}(1), \mathbf{z}(2), \cdots, \mathbf{z}(m))$ based on the training set, the "missing-view-in-test-data" problem requires a classifier in the form of $\tilde{f}_{\tilde{\boldsymbol{\theta}}} : \mathcal{X}_1 \to \mathcal{L}$ due to the missing auxiliary view in the test data. To effectively incorporate the available information in the entire training set $\{\mathbf{z}(i), l(i)\}_{i=1}^m$, this paper focuses on constructing the classifier $\tilde{f}_{\tilde{\boldsymbol{\theta}}}$, where $\tilde{\boldsymbol{\theta}} = \tilde{\boldsymbol{\theta}}(\mathbf{z}(1), \mathbf{z}(2), \cdots, \mathbf{z}(m), l(1), l(2), \cdots, l(m))$. To capture the information hidden in the auxiliary view $\mathbf{x}_2(j)$ and training labels $l(j)$, an intermediate space $\mathcal{S}$ is constructed to maximally retain the discriminative information from both views.

During this process, we construct projections $p_1(\cdot)$ and $p_2(\cdot)$ that map $\{\mathbf{x}_1(i)\}_{i=1}^m$ and $\{\mathbf{x}_2(i)\}_{i=1}^m$ to $p_1(\mathbf{x}_1(i)) \in \mathcal{S}$ and $p_2(\mathbf{x}_2(i)) \in \mathcal{S}$, respectively. In $\mathcal{S}$, the classification problem becomes easier: a discriminative classifier $f : \mathcal{S} \to \mathcal{L}$ (*e.g.*, the SVM classifier [22]) is trained based on $\{p_1(\mathbf{x}_1(i)), l(i)\}_{i=1}^m \cup \{p_2(\mathbf{x}_2(i)), l(i)\}_{i=1}^m$. The training process is shown in blue arrows in Fig. 1(a). In the test phase, the test samples $\{\hat{\mathbf{x}}_1(j)\}_{j=1}^k$ are first projected to $p_1(\hat{\mathbf{x}}(j)) \in \mathcal{S}$ and subsequently fed to the trained classifier $f : \mathcal{S} \to \mathcal{L}$. The test process is shown in yellow arrows in Fig. 1(a).

The success of the aforementioned latent space based approach depends on not only the maximal preservation of the congruent information among views, but also the discriminative label information acquired in constructing $\mathcal{S}$. To achieve this, we propose a discriminative canonical correlation analysis (DCCA) algorithm, which simultaneously extracts multiple CCA projections and also incorporates the label information. In the following section, we formally formulate the optimization problem of DCCA and compare our proposed optimization method against previous ones.

4 DCCA: Formulation

In this section, we formulate the DCCA algorithm and compare it with previous related work. The classical Canonical Correlation Analysis (CCA) (see [23, 24, 11]) and its variants have been popular among practitioners for decades. In its basic form, the CCA algorithm finds a pair of linear projection vectors that maximize the linear correlation:

$$\begin{aligned} \max_{\boldsymbol{\alpha}_j,\boldsymbol{\alpha}_k} \quad & \boldsymbol{\alpha}_j^T \Sigma_{jk} \boldsymbol{\alpha}_k \\ \text{s.t.} \quad & \boldsymbol{\alpha}_j^T \Sigma_{jj} \boldsymbol{\alpha}_j = 1 \quad (j,k = 1,2,\cdots,J) \end{aligned} \tag{1}$$

where $\Sigma_{jk} = \mathcal{E}(\mathbf{x}_j \mathbf{x}_k^T)$, $\mathcal{E}$ denotes the mathematical expectation, and $\mathbf{x}_j$ denotes the observations in the jth view ($j = 1,2,\cdots,J$, J denotes the number of distinct views in the multi-view data). In this way, $\boldsymbol{\alpha}_j$ and $\boldsymbol{\alpha}_k$ project the jth and kth views to a one-dimensional space by computing the inner products $\boldsymbol{\alpha}_j^T \hat{\mathbf{x}}_j$ and $\boldsymbol{\alpha}_k^T \hat{\mathbf{x}}_k$, respectively. In practice, a latent space of more of dimension d ($d > 1$) is often desired, hence this procedure in Eq. (1) needs to be repeated d times. Alternatively, as shown in [11], the following formulation can be used equivalently:

$$\begin{aligned} \max_{A_j,A_k} \quad & \mathrm{tr}(A_j^T \Sigma_{jk} A_k) \\ \text{s.t.} \quad & A_j^T \Sigma_{jj} A_j = \mathbf{I} \quad (j,k = 1,2,\cdots,J) \end{aligned} \tag{2}$$

where $A_j = [\boldsymbol{\alpha}_j(1),\cdots,\boldsymbol{\alpha}_j(d)]$ contains all d projection vectors to form a d-dimensional latent space. The projections are computed as $A_j^T \hat{\mathbf{x}}_j$ and $A_k^T \hat{\mathbf{x}}_k$. To preserve the semantic consistency between the jth and kth views, we strive to make sure that $A_j^T \hat{\mathbf{x}}_j$ and $A_k^T \hat{\mathbf{x}}_k$ are as similar as possible, therefore maximizing $\mathrm{tr}(A_j^T \Sigma_{jk} A_k)$ as Eq. (2).

These CCA algorithms are extended to handle more than two views, which are generally termed as the generalized CCA (gCCA) [11, 25]. However, among the gCCA algorithms where per-view "correlations" are combined nonlinearly, most of them cannot be solved in a closed form and require a computing scheme to optimize the projections one by one. Also, the naive version of gCCA ignores the discriminative information in the class labels provided in the training process.

To address these two potential issues, we propose a new algorithm, named as *Discriminative CCA* (DCCA), which is inspired by [10]. It not only inherits the guaranteed optimization convergence property, but also has two extended properties: 1) the simultaneous optimization of multiple CCA projections contributes to better projection coefficients even in a kernelized version; 2) the pursuit of DCCA incorporates the label view information. Empirically, we find that the combination of 1) and 2) leads to performance gains across three challenging visual datasets, as shown in Section 6.

Starting from leveraging discriminative information like Loog *et al.* [26], we first encode the training labels as the Jth view based on k-nearest neighbors (k-NNs) of the training samples, where observations from all views are normalized and concatenated and then used in computing k-NNs. Suppose that there are C classes. For each labeled training sample $\mathbf{x}(0) \overset{\text{def}}{=} [\mathbf{x}_1(0)^T, \cdots, \mathbf{x}_J(0)^T]^T$, we consider its k-NNs $\mathbf{x}(i) \overset{\text{def}}{=} [\mathbf{x}_1(i)^T, \cdots, \mathbf{x}_J(i)^T]^T$ $(i = 0, \cdots, k$, including itself $\mathbf{x}(0))$. We record their labels as $\mathbf{L}_{(i)}$ $(i = 0, \cdots, k)$ accordingly. The encoded label view for this training sample is a binary vector consisting of $k+1$ length-C blocks.

After obtaining the label view, we can simultaneously optimize all d projections in the form of A_j $(j = 1, 2, \cdots, J)$ as follows:

$$\begin{aligned} &\arg\max_{\mathbf{A}} \sum_{j,k=1,j\neq k}^{J} c_{jk} g\left(\operatorname{tr}(A_j^T \Sigma_{jk} A_k)\right) \\ &\text{s.t.} \qquad A_j^T \Sigma_{jj} A_j = \mathbf{I} \quad (j = 1, 2, \cdots, J) \end{aligned} \tag{3}$$

where $\mathbf{A} \overset{\text{def}}{=} [A_1, \cdots, A_J]$, $A_j \overset{\text{def}}{=} [\boldsymbol{\alpha}_j(1), \cdots, \boldsymbol{\alpha}_j(d)]$ projects $\mathbf{x}_j$ to the d-dimensional common semantic space, c_{jk} is the selecting weights (either 0 or 1), and function $g(\cdot)$ is the view combination function (*e.g.*, direct combination $g(x) = x$, or squared combination $g(x) = x^2$). The space dimension d is chosen such that each diagonal element of $A_j^T \Sigma_{jk} A_k$ is positive, which implies that the observations from all views are positively correlated. The Lagrangian of Eq. (3) is

$$F(\mathbf{A}, \mathbf{\Lambda}) = \sum_{k\neq j} c_{jk} g\left(\operatorname{tr}(A_j^T \Sigma_{jk} A_k)\right) - \phi \sum_j \frac{1}{2} \operatorname{tr}\left(\mathbf{\Lambda}_j^T (A_j^T \Sigma_{jj} A_j - \mathbf{I})\right) \tag{4}$$

where $\mathbf{\Lambda} \overset{\text{def}}{=} [\mathbf{\Lambda}_1, \cdots, \mathbf{\Lambda}_J]$, $\mathbf{\Lambda}_j \in \mathcal{R}^{d\times d}$ is the multiplier, and ϕ is a scalar which is equal to 1 if $g(x) = x$ and 2 if $g(x) = x^2$. The derivative of $g(x)$ is denoted as $g'(x)$.

From Eq. (4), the following stationary equations hold:

$$\frac{1}{\phi} \sum_{k\neq j} c_{jk} g'\left(\operatorname{tr}(A_j^T \Sigma_{jk} A_k)\right) \Sigma_{jk} A_k = \Sigma_{jj} A_j \mathbf{\Lambda}_j;\ A_j^T \Sigma_{jj} A_j = \mathbf{I} \tag{5}$$

In practice, a kernelized version of DCCA is often favored, because linear correlations are not sufficient in modeling the nonlinear interplay among different views. Suppose that $\boldsymbol{K}_j$ is the Gram matrix of the centered data points $[\mathbf{x}_1, \cdots, \mathbf{x}_J]$. The empirical covariance is $\frac{1}{n} \boldsymbol{\alpha}_j^T \boldsymbol{K}_j \boldsymbol{K}_k \boldsymbol{\alpha}_k$, where $\boldsymbol{\alpha}_j$ is the coefficient vector and n is the number of training samples. Let $A_j = [\boldsymbol{\alpha}_j(1), \cdots, \boldsymbol{\alpha}_j(d)]$ and the empirical covariance matrix be $\frac{1}{n} A_j^T \boldsymbol{K}_j \boldsymbol{K}_k A_k$. Suppose that $\boldsymbol{K}_j$ can be decomposed as $\boldsymbol{K}_j = \boldsymbol{R}_j^T \boldsymbol{R}_j$. We define the projection matrix $\mathbf{W}_j = \boldsymbol{R}_j A_j$, and similarly define $\mathcal{W} \overset{\text{def}}{=} [\mathbf{W}_1, \cdots, \mathbf{W}_J]$. The optimization objective function of the kernelized DCCA is

$$\begin{aligned} \arg\max_{\mathcal{W}} \quad & \sum_{j,k=1,j\neq k}^{J} c_{jk} g\left(\operatorname{tr}(\frac{1}{n}\mathbf{W}_j^T \boldsymbol{R}_j \boldsymbol{R}_k^T \mathbf{W}_k)\right) \\ \text{s.t.} \quad & \mathbf{W}_j^T \left[(1-\tau_j)\frac{1}{n}\boldsymbol{R}_j \boldsymbol{R}_j^T + \tau_j \mathbf{I}_n\right] \mathbf{W}_j = \mathbf{I} \end{aligned} \tag{6}$$

Let us define $N_j = (1-\tau_j)\frac{1}{n}\boldsymbol{R}_j\boldsymbol{R}_j^T + \tau_j\mathbf{I}_n$, where $0 < \tau < 1$ is a pre-specified regularization parameter. Similar to Eq. (5), the following stationary equations hold

$$\frac{1}{\phi}\sum_{k\neq j} c_{jk} g'\left(\operatorname{tr}(\frac{1}{n}\mathbf{W}_j^T \boldsymbol{R}_j \boldsymbol{R}_k^T \mathbf{W}_k)\right)\frac{1}{n}\boldsymbol{R}_j \boldsymbol{R}_k^T \mathbf{W}_k = N_j \mathbf{W}_j \boldsymbol{\Lambda}_j \tag{7}$$

$$\mathbf{W}_j^T N_j \mathbf{W}_j = \mathbf{I} \tag{8}$$

whose solution[1] is presented in Section 5.

5 DCCA: Solution

In this section, a monotonically convergent iterative algorithm to solve the D-CCA problem is presented. For generic $g(\cdot)$ and c_{jk} value assignments, there is no closed-form solution to Eq. (5) or Eq. (7). However, following [10], a similar "PLS-type", monotonically convergent iterative algorithm can be formulated. For conciseness, we only present the details of this algorithm with the solution to the problem in Eq. (7). Define the outer component $\mathbf{Y}_j$ and inner component $\mathbf{Z}_j$, respectively, as

$$\mathbf{Y}_j \overset{\text{def}}{=} \boldsymbol{R}_j^T \mathbf{W}_j \tag{9}$$

$$\mathbf{Z}_j \overset{\text{def}}{=} \frac{1}{\phi}\sum_{k\neq j} c_{jk} g'\left(\operatorname{tr}(\frac{1}{n}\mathbf{W}_j^T \boldsymbol{R}_j \boldsymbol{R}_k^T \mathbf{W}_k)\right)\mathbf{Y}_k \tag{10}$$

Differentiating the Lagrangian with respect to $\mathbf{W}_j$ and setting the gradient to zero, we obtain

$$\boldsymbol{R}_j \mathbf{Z}_j = N_j \mathbf{W}_j \boldsymbol{\Lambda}_j;\ \mathbf{W}_j^T N_j \mathbf{W}_j = \mathbf{I} \tag{11}$$

From Eq. (11) and Eq. (10), we have

$$\boldsymbol{\Lambda}_j = \mathbf{W}_j^T \boldsymbol{R}_j \mathbf{Z}_j = \frac{1}{\phi n}\sum_{k\neq j} c_{jk} g'\left(\operatorname{tr}(\frac{1}{n}\mathbf{W}_j^T \boldsymbol{R}_j \boldsymbol{R}_k^T \mathbf{W}_k)\right)\mathbf{W}_j^T \boldsymbol{R}_j \boldsymbol{R}_k^T \mathbf{W}_k \tag{12}$$

where $\operatorname{tr}(\frac{1}{n}\mathbf{W}_j^T \boldsymbol{R}_j \boldsymbol{R}_k^T \mathbf{W}_k)$ is assumed to be positive, $c_{jk} = 0$ or 1, and due to the definition of g', $\boldsymbol{\Lambda}_j$ is a positive semi-definite matrix. From Eq. (11) and Eq. (12),

[1] Note that Eq. (5) is a linear version of Eq. (7) and has a very similar solution. For conciseness, the solution to Eq. (5) is omitted.

we have $\mathbf{\Lambda}_j^T\mathbf{\Lambda}_j = \mathbf{Z}_j^T\boldsymbol{R}_j^T N_j^{-1}\boldsymbol{R}_j\mathbf{Z}_j$. Since $\mathbf{\Lambda}_j$ has non-negative eigenvalues, $\mathbf{\Lambda}_j$ can be obtained via the matrix square root $[\mathbf{Z}_j^T\boldsymbol{R}_j^T N_j^{-1}\boldsymbol{R}_j\mathbf{Z}_j]^{1/2}$. Therefore,

$$\mathbf{W}_j = N_j^{-1}\boldsymbol{R}_j\mathbf{Z}_j\left(\left[\mathbf{Z}_j^T\boldsymbol{R}_j^T N_j^{-1}\boldsymbol{R}_j\mathbf{Z}_j\right]^{1/2}\right)^{\dagger} \tag{13}$$

where † denotes the pseudoinverse of a matrix.

A monotonically convergent iterative algorithm is described in Algorithm 1. Let $f(\mathcal{W})\overset{\text{def}}{=}\sum_{k\neq j} c_{jk}\, g\left(\text{tr}(\frac{1}{n}\mathbf{W}_j^T\boldsymbol{R}_j\boldsymbol{R}_k^T\mathbf{W}_k)\right)$. Importantly, we have the following proposition:

Proposition 1. $f(\mathcal{W}(s=1)) \leq f(\mathcal{W}(s=2)) \leq f(\mathcal{W}(s=3)) \leq \ldots \leq C_u < \infty$ *holds for all* $s \in \mathcal{N}$*, where* s *denotes the iteration index and* C_u *is a constant bound. This guarantees Algorithm 1 to converge monotonically.*

Proof. Due to space limit, we defer the proof to the supplemental material. □

Algorithm 1: A monotonically convergent iterative algorithm for DCCA

```
Input: Observations from all $J$ views: $\mathbf{x}_j, j = 1, \cdots, J$.
Output: $J$ projection matrices $\mathbf{W}_j, j = 1, \cdots, J$.
Initialization:
Randomly initialize $\mathbf{W}_j(0)$, normalize them by
$\mathbf{W}_j(0) \leftarrow N_j^{-1}\mathbf{W}_j(0)\left(\left[\mathbf{W}_j(0)^T N_j^{-1}\mathbf{W}_j(0)\right]^{1/2}\right)^{\dagger}$, and compute the initial outer
components $\mathbf{Y}_j(0)$ by Eq. (9).
for $s = 0, 1, \cdots$ until the convergence of $\mathbf{W}_j$ do
    for $j = 1, 2, \cdots J$ do
        update the inner components by Eq. (10):
        $\mathbf{Z}_j(s) \leftarrow \frac{1}{\phi}\sum_{k=1}^{j-1} c_{jk} g'(\text{tr}(\frac{1}{n}\mathbf{Y}_j^T(s)\mathbf{Y}_k(s+1)))\mathbf{Y}_k(s+1) +$
        $\frac{1}{\phi}\sum_{k=j+1}^{J} c_{jk} g'(\text{tr}(\frac{1}{n}\mathbf{Y}_j^T(s)\mathbf{Y}_k(s)))\mathbf{Y}_k(s)$;
        update the outer weighs by Eq. (13):
        $\mathbf{W}_j(s+1) \leftarrow N_j^{-1}\boldsymbol{R}_j\mathbf{Z}_j(s)\left(\left[\mathbf{Z}_j(s)^T\boldsymbol{R}_j^T N_j^{-1}\boldsymbol{R}_j\mathbf{Z}_j(s)\right]^{1/2}\right)^{\dagger}$;
        update the outer components by Eq. (9): $\mathbf{Y}_j(s+1) \leftarrow \boldsymbol{R}_j^T\mathbf{W}_j(s+1)$;
    end
end
return
```

6 Experiments

6.1 Compared Methods and Datasets

We compare the performances of the kernelized DCCA algorithm (paired with the RBF SVM classifier) against the following algorithms: A vanilla “SVM”:

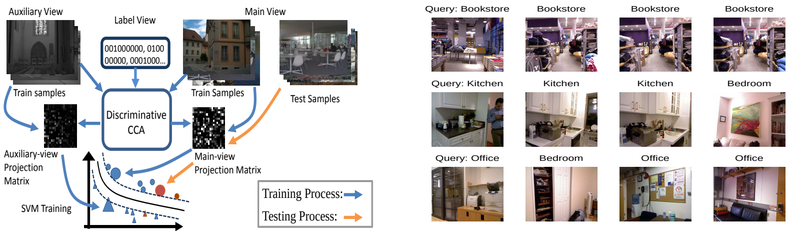
that is trained on the main view of the visual information only. A variant of "SVM2K" [5]: which we train with two views in the train phase but only use one of the classifier due to the missing auxiliary view in the test phase[2]. The "KCCA": a kernel CCA based on the main view and the auxiliary view. The "KCCA+L": a kernel CCA based on the main view and the encoded label view, it ignores the auxiliary view. The "RGCCA": a kernel variant of the regularized generalized CCA [10] based on the main view and the auxiliary view. As an extended version of gCCA, it iteratively optimizes each one-dimensional projection vectors (one column of $\mathbf{W}_j$), and all d columns of $\mathbf{W}_j$ are pursued one by one, similar to the algorithm presented in [24]. The "RGCCA+L": similar to "RGCCA", except it is based on the main view and the encoded label view. The "RGCCA+AL": similar to the proposed DCCA, except it iteratively optimizes each one-dimensional projection vectors (one column of $\mathbf{W}_j$).

In selecting the competing algorithms, we choose the "SVM2K" to represent a variant of multi-view learning algorithm. We select "KCCA" and "RGCCA" to represent classical and recent implementations of the CCA algorithms, both of which ignore the encoded label view. To isolate the effects of the auxiliary view and encoded train label view, we have included a series of combinations "KCCA+L", "RGCCA+L" and "RGCCA+AL". In the following experiments, the RBF kernel is applied in both the KCCA algorithm and the SVM algorithms. Parameters such as c_{jk}, $g(\cdot)$, the bandwidth parameter in the RBF kernel, and those parameters in the SVM2K algorithm, are all selected by a 4-fold cross validation. The experiments are conducted on three different datasets, *i.e.*, the "NYU Depth V1" Indoor Scenes dataset ([27]), the RGBD Object dataset [28], and the multi-spectral scene dataset [29]. The NYU Depth V1 dataset consists of RGBD images of indoor scenes collected by a modified Kinect sensor [27]. With this dataset, we demonstrate that the depth information in the train phase can benefit the scene classification based solely on the RGB images. The RGBD object dataset [28] consists of a large collection of paired RGB images and depth maps of common objects. We focus on the instance level object recognition, and demonstrate that the additional depth information during the train phase can facilitate better recognition based on the RGB information only. The multi-spectral scene dataset [29] consists of 477 registered and aligned RGB and near-infrared (IR) images from 9 different scene categories, *i.e.*, country, field, forest, indoor, mountain, old-building, street, urban and water. In this experiment, we demonstrate that the auxiliary information hidden in the IR channel can help to train a better scene recognition model that operates only on the main view.

6.2 NYU-Depth-V1-Indoor Scene Dataset

On the NYU Depth V1 indoor scenes dataset [27], we carry out the multi-spectral scene recognition task. Following [27], the scene observations are randomly split into 10 folds with equal size of the train set and the test set. Subsequently, we

[2] The original form of SVM2K is not directly applicable to the missing view problem



(a) (b)

Fig. 1: (a) A latent space model to address the "missing-view-in-test-data" problem. The training and test processes are displayed in blue and yellow arrows, respectively. (b) k-NN retrieval from the NYU depth Indoor scenes data base. Actual scene class labels are listed above each image.

Table 1: NYU Depth V1 Indoor Scenes Classification, the highest and second highest values are colored red and blue, respectively.

Features	GIST		SP200		SP800	
Views	L+D	RGB+D	L+D	RGB+D	L+D	RGB+D
SVM	59.57±3.31	60.79±3.12	64.11±3.11	64.71±3.95	64.73±2.79	65.34±3.18
SVM2K	57.52±3.88	60.01±3.71	59.62±3.23	60.42±4.55	58.00±3.58	60.64±3.84
KCCA	58.16±6.55	62.58±3.55	64.94±4.58	64.00±4.92	64.77±4.69	65.01±4.86
KCCA+L	58.48±3.37	59.95±3.62	62.99±3.80	60.67±4.23	62.26±3.56	60.55±4.40
RGCCA	58.66±5.93	59.75±4.11	60.49±5.21	60.31±5.75	61.70±4.00	60.42±3.68
RGCCA+L	59.12±4.11	59.82±4.50	63.34±4.18	62.48±3.49	63.81±4.51	61.04±4.99
RGCCA+AL	**59.82**±6.10	**62.85**±4.24	**65.61**±4.22	**65.31**±4.23	**65.38**±4.22	**65.66**±3.04
DCCA	**60.26**±3.86	**63.60**±3.43	**66.20**±3.69	**65.35**±4.72	**66.09**±4.18	**66.28**±4.16

extract both the GIST [30] features and the spatial pyramid bag-of-feature representation [31] independently from each channel. For the latter, we densely extract SIFT descriptors from 40 × 40 patches with a stride of 10 pixels, and use two k-means dictionaries of sizes 200 and 800 to build the representation, which are shorted as SP200 and SP800, respectively. While grouping the imaging channels into views, we investigate the following two settings: (1) **L+D**: Grayscale image features are assigned as the main view, while the depth features are assigned as the auxiliary view. (2) **RGB+D**: RGB image features are concatenated and assigned as the main view, while the depth features are assigned as the auxiliary view.

We first demonstrate the k-NN retrieval results in Fig. 1(b) to visualize some typical images from this dataset. The query images (from the test set) are displayed on the left and the corresponding 3 nearest neighbors (in the train set) are displayed on the right. Clearly, this dataset consists of highly cluttered indoor scenes, making this scene recognition task a challenging one.

In Table 1, the means and standard deviations of the recognition accuracy (in percentage) are reported. we observe that higher dimensional features offer better accuracy and the color information also helps recognition slightly. Generally, experiments based on "RGB+D" features achieve slightly higher accuracies than their "L+D" counterparts.

In Table 1, the SVM2K-variant gives lower accuracies than the SVM baseline, we speculate that the loose "prediction consistency" regularization of SVM2K-variant is only helpful when two views satisfy certain distributions. In addition, neither the KCCA nor the RGCCA approach sees significant performance improvements.

With the Label view alone, neither the "KCCA+L" nor the "RGCCA+L" algorithm achieves any advantage over the baseline. Intuitively, the information embedded in the label view is far less significant than that in the auxiliary view. However, with both the label view and the auxiliary view, the "RGCCA+AL" algorithm is capable of achieving a small advantage over the baseline, though not as significant as the proposed DCCA algorithm, whose projections are optimized and computed simultaneously and more accurately.

The large standard deviation in Table 1 stem from the difficult nature of the datasets, which can be seen in the baseline SVM performance in Table 1.

6.3 RGBD Object Dataset

With this RGBD Object dataset from [28], we focus on the instance level object recognition. There are multiple instances of objects across all the 51 categories, the target is to correctly recover the object instance labels in the test set. We follow the "leave-one-sequence-out" scheme in [28] and split recordings with camera mounting angle of 30° and 60° as the train set and the remaining recordings as the test set (the train/test sets are fixed, hence the standard deviation are not defined). In this section, we present the results with the EMK-based features [28] extracted from the RGB channels as the main view, and depth channel as the auxiliary view. In addition, we have also verified the efficacy of the proposed method with the state-of-the-art HMP-based features [32], but we defer the detailed results and comments to the supplemental material due to limited space.

As is seen in Table 2–3(too many entries to fit in a single page), the recognition accuracy (in percentage) within each category fluctuates significantly. With some of the easy categories (*e.g.*, "pitcher"), the baseline SVM algorithm already achieves perfect recognition. However, with some of the challenging categories (*e.g.*, "food bag" and "lime"), the proposed DCCA offers the most significant performance boost. Overall, the "KCCA+L" and the "RGCCA+L" algorithms achieve a small advantage over the SVM baseline, both of which are inferior to the RGCCA algorithm that only maximizes the main view and auxiliary view correlation. However, the "RGCCA+AL" algorithm performs much better, though not as good as the proposed DCCA algorithm. Among the 51 categories in Table 2–3, the "RGCCA+AL" algorithm achieves the best and second best accuracies in 8 and 24 categories, earning itself an overall average accuracy of

85.7%. The proposed DCCA achieves the best and second best recognition accuracies in 28 and 19 categories, acquiring an average accuracy of 86.6% across all categories, highest among all algorithms.

6.4 Multi-Spectral Scene Dataset

Following [29], we construct 10 random training/testing splits. For each split, 99 RGB images (11 per category) are used as the test set while the remaining 378 pairs as the train set. Before the feature extraction, each RGB image is converted to the **LAB** color space (similarly to [29]). Then the GIST [30] features are computed independently on each of the channels. Of these four channels, we choose the following two view assignment schemes: (1) **L+I**: Grayscale GIST features and the IR channel GIST features are assigned as the main view and the auxiliary view, respectively. (2) **LAB+I**: GIST features extracted independently from the L,A and B channels are concatenated as the main view, while those extracted from the IR channel are considered as the auxiliary view. In Table 4, the mean and the standard deviation (both in percentage) of the recognition accuracies are reported. We observe that under either view assignment, neither the KCCA nor the SVM2K algorithm achieves a significant advantage over the SVM baseline. With the label view alone, "KCCA+L" and "RGCCA+L" achieve recognition accuracy on a par with the baseline SVM. We speculate that the auxiliary view is more informative in this dataset: with the auxiliary view, the RGCCA algorithm is capable of outperform the baseline by a small margin. Furthermore, with the additional label view information in "RGCCA+AL", this margin is enlarged. Overall, the proposed DCCA still outperforms all other competing algorithms. The large standard deviations in Table 4 could stem from the nature of this dataset. Indeed, in [29], Brown and Susstrunk also report large standard deviations on a par with ours.

6.5 Discussion

Overall, based on the aforementioned empirical results, we have the following observations. The latent space-based model is capable of leveraging the information from the auxiliary view in training, and hence effectively address the missing-view-in-test-data problem. Without the auxiliary view, the encoded label view alone is not significant enough to evidently boost the recognition performance. Incorporating the encoded label view with the auxiliary view yields some additional boost in recognition performance. DCCA consists of three components: the incorporation of the auxiliary view, the encoded label view, and the simultaneous optimization. They jointly contribute to the performance gains.

7 Conclusions

In this paper, we explored a practical visual recognition problem, where we have multi-view data in the train phase but only the single-view data in the test phase.

Table 2: Accuracy Table Part 1 for the Multi-View RGBD Object Instance recognition, the highest and second highest values are colored red and blue, respectively. The remaining part is in Table 3.

Category	SVM	SVM2K	KCCA	KCCA	RGCCA	RGCCA	RGCCA	DCCA
				+L		+L	+AL	
apple	65.2	72.4	**77.6**	64.8	**79.0**	68.1	76.7	**77.6**
ball	95.9	97.3	97.8	**99.5**	94.2	95.3	97.8	**98.4**
banana	74.2	61.1	**80.3**	71.7	**77.3**	**77.3**	**80.3**	**80.3**
bell pepper	**71.3**	59.8	**69.3**	65.7	66.1	68.9	**69.3**	**69.3**
binder	67.3	36.7	74.1	52.4	**75.5**	68.7	74.1	**74.8**
bowl	85.0	85.8	87.7	81.2	**90.0**	86.2	**88.1**	**88.1**
calculator	**99.4**	88.3	**99.4**	**100**	97.8	**99.4**	**99.4**	**99.4**
camera	91.7	49.6	**97.5**	90.1	**96.7**	95.9	**97.5**	**97.5**
cap	91.8	88.9	**95.9**	91.2	**96.5**	93.0	**95.9**	**95.9**
cellphone	93.2	81.7	**96.3**	93.2	94.2	**95.3**	**96.3**	**96.3**
cereal box	77.4	75.7	**82.5**	**89.8**	**82.5**	80.8	81.9	**82.5**
coffee mug	82.7	81.7	**89.2**	62.5	81.7	82.7	**87.3**	**89.2**
comb	97.3	96.0	**99.3**	**100**	98.7	98.7	**99.3**	**100**
dry battery	**90.3**	79.3	86.3	87.7	**92.5**	86.3	88.1	87.7
flashlight	77.1	75.0	80.3	74.5	78.2	77.7	**81.4**	**82.4**
food bag	72.7	69.1	80.8	82.5	84.9	77.7	**86.6**	**89.2**
food box	75.1	72.6	78.0	83.8	84.1	76.1	**84.4**	**86.4**
food can	70.0	63.7	70.7	78.2	73.7	66.2	**81.0**	**83.7**
food cup	87.1	86.4	84.9	**94.5**	83.8	84.2	90.1	**91.5**
food jar	84.5	81.0	**88.6**	86.1	85.8	85.4	88.3	**88.9**
garlic	**95.5**	**93.3**	92.2	**95.5**	89.0	91.2	92.8	**93.3**
glue stick	**100**	89.3	99.4	95.6	**100**	93.7	**99.7**	99.4
greens	75.7	70.3	**82.2**	**84.9**	74.6	80.5	81.1	**82.2**
hand towel	80.1	74.5	80.9	82.4	79.0	78.7	**83.9**	**85.4**
instant noodles	83.1	78.7	**97.1**	88.8	89.5	88.5	**95.4**	**97.1**
keyboard	88.1	82.7	90.6	88.1	**93.6**	88.1	**95.0**	**95.0**
kleenex	**96.6**	90.9	95.1	89.8	94.7	93.6	**95.8**	**95.8**
lemon	45.8	44.2	43.4	45.0	44.2	43.0	**51.8**	**53.0**
light bulb	93.2	92.5	**95.9**	**98.6**	**95.9**	95.2	95.2	**95.9**
lime	37.8	38.3	42.2	37.2	45.0	40.0	**48.3**	**50.0**
marker	**48.4**	39.3	46.0	**48.4**	**49.9**	46.2	46.6	46.9
mushroom	**100**	**99.4**	**100**	**100**	**100**	**99.4**	**100**	**100**
notebook	80.5	73.3	**86.5**	82.0	**85.7**	82.7	**85.7**	**86.5**
onion	91.5	86.4	88.6	**94.0**	93.1	88.6	**93.4**	**93.4**
orange	41.1	42.0	**57.0**	49.8	19.3	48.3	**51.7**	**57.0**
peach	**100**	74.2	**100**	97.4	**99.3**	98.7	**100**	**100**
pear	68.7	61.6	79.0	70.1	79.0	74.0	**81.9**	**84.0**
pitcher	**100**	**100**	**100**	**100**	**100**	**98.3**	**100**	**100**
plate	89.8	74.2	96.3	85.8	96.9	91.9	**97.6**	**98.0**
pliers	78.2	62.4	**86.5**	79.5	72.1	80.8	84.7	**87.3**
potato	62.9	65.6	**70.3**	**71.8**	64.5	67.6	69.1	**70.3**
rubber eraser	**99.0**	75.5	95.1	96.6	93.1	96.1	97.5	**98.5**
scissors	87.5	84.9	**96.7**	**96.7**	**96.1**	92.8	**96.1**	**96.7**
shampoo	84.5	82.9	94.5	82.3	**95.5**	89.0	94.2	**94.8**
soda can	89.6	87.8	91.0	**97.7**	92.8	88.7	95.5	**96.8**

Table 3: Continued from Table 2: Accuracy Table Part 2 for the Multi-View RGBD Object Instance recognition, the highest and second highest values are colored red and blue, respectively.

Category	SVM	SVM2K	KCCA	KCCA+L	RGCCA	RGCCA+L	RGCCA+AL	DCCA
sponge	76.4	64.8	75.2	69.4	**78.6**	72.8	76.4	**77.4**
stapler	71.8	67.7	73.3	70.6	74.2	71.5	**77.4**	**78.0**
tomato	81.9	70.0	79.6	76.8	**82.2**	77.1	**85.0**	**85.0**
tooth brush	**78.4**	69.6	76.8	70.1	**79.4**	74.7	77.8	77.8
tooth paste	84.5	68.3	**90.0**	86.0	81.9	87.1	88.9	**90.4**
water bottle	88.2	88.2	**90.6**	82.6	80.2	87.4	**89.3**	**90.6**
average	81.3	74.4	84.5	81.6	83.0	81.8	**85.7**	**86.6**

Table 4: Multi-Spectral Scene recognition, the highest and second highest values are colored red and blue, respectively.

Views	SVM	SVM2K	KCCA	KCCA+L
LAB+I	67.78±5.25	67.17±5.58	66.87±4.76	66.46±2.83
L+I	61.82±3.77	61.82±4.59	62.32±4.81	61.92±3.64

Views	RGCCA	RGCCA+L	RGCCA+AL	DCCA
LAB+I	68.59±3.94	67.27±4.88	**69.90**±3.16	**70.51**±2.37
L+I	62.22±3.96	62.42±4.45	**64.55**±3.44	**65.66**±4.57

We have verified that information from the auxiliary view in the train data can indeed lead to better recognition in the test phase even when the auxiliary view is entirely missing. As a part of our verification-by-construction proof, we have proposed a new discriminative canonical correlation analysis to integrate and map the semantic information from all views to a common latent space, over which all subsequent classification is conducted. We also investigated and isolated the effects of the encoded label view and the auxiliary view. The experimental results demonstrate that the proposed approach achieves performance advantages on all three benchmarks.

Acknowledgement. Research reported in this publication was partly supported by the National Institute Of Nursing Research of the National Institutes of Health under Award Number R01NR015371. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health. This work is also partly supported by US National Science Foundation Grant IIS 1350763, China National Natural Science Foundation Grant 61228303, GH's start-up funds form Stevens Institute of Technology, a Google Research Faculty Award, a gift grant from Microsoft Research, and a gift grant from NEC Labs America.

References

1. Quanz, B., Huan, J.: Large margin transductive transfer learning. In: Proceedings of the 18th ACM conference on Information and knowledge management, ACM (2009) 1327–1336
2. Davis, J.V., Kulis, B., Jain, P., Sra, S., Dhillon, I.S.: Information-theoretic metric learning. In: Proceedings of the 24th international conference on Machine learning, ACM (2007) 209–216
3. Saenko, K., Kulis, B., Fritz, M., Darrell, T.: Adapting visual category models to new domains. In: Computer Vision–ECCV 2010. Springer (2010) 213–226
4. Kulis, B., Saenko, K., Darrell, T.: What you saw is not what you get: Domain adaptation using asymmetric kernel transforms. In: Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on, IEEE (2011) 1785–1792
5. Farquhar, J., Hardoon, D., Meng, H., Shawe-taylor, J.S., Szedmak, S.: Two view learning: Svm-2k, theory and practice. In: Advances in neural information processing systems. (2005) 355–362
6. Zhang, D., He, J., Liu, Y., Si, L., Lawrence, R.D.: Multi-view transfer learning with a large margin approach. In: Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining. (2011) 1208–1216
7. Qi, Z., Yang, M., Zhang, Z.M., Zhang, Z.: Mining noisy tagging from multi-label space. In: Proceedings of the 21st ACM international conference on Information and knowledge management, ACM (2012) 1925–1929
8. Vapnik, V., Vashist, A., Pavlovitch, N.: Learning using hidden information (learning with teacher). In: Neural Networks, 2009. IJCNN 2009. International Joint Conference on, IEEE (2009) 3188–3195
9. Argyriou, A., Evgeniou, T., Pontil, M., Argyriou, A., Evgeniou, T., Pontil, M.: Convex multi-task feature learning. In: Machine Learning, press (2007)
10. Tenenhaus, A., Tenenhaus, M.: Regularized generalized canonical correlation analysis. Psychometrika **76** (2011) 257–284
11. Hardoon, D.R., Szedmak, S., Shawe-Taylor, J.: Canonical correlation analysis: An overview with application to learning methods. Neural Computation **16** (2004) 2639–2664
12. Kulis, B., Sustik, M., Dhillon, I.: Learning low-rank kernel matrices. In: Proceedings of the 23rd international conference on Machine learning, ACM (2006) 505–512
13. Blum, A., Mitchell, T.: Combining labeled and unlabeled data with co-training. In: Proceedings of the eleventh annual conference on Computational learning theory, ACM (1998) 92–100
14. Chen, L., Li, W., Xu, D.: Recognizing rgb images by learning from rgb-d data. In: Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on, IEEE (2014)
15. Shrivastava, A., Gupta, A.: Building part-based object detectors via 3d geometry. In: Computer Vision (ICCV), 2013 IEEE International Conference on, IEEE (2013) 1745–1752
16. Tommasi, T., Quadrianto, N., Caputo, B., Lampert, C.H.: Beyond dataset bias: Multi-task unaligned shared knowledge transfer. In: Computer Vision–ACCV 2012. Springer (2013) 1–15
17. Globerson, A., Roweis, S.: Nightmare at test time: robust learning by feature deletion. In: Proceedings of the 23rd international conference on Machine learning, ACM (2006) 353–360

18. Srivastava, N., Salakhutdinov, R.: Multimodal learning with deep boltzmann machines. In: Advances in neural information processing systems. (2012) 2222–2230
19. Chen, J., Liu, X., Lyu, S.: Boosting with side information. In: Computer Vision–ACCV 2012. Springer (2013) 563–577
20. Shams, L., Wozny, D.R., Kim, R., Seitz, A.: Influences of multisensory experience on subsequent unisensory processing. Frontiers in psychology **2** (2011)
21. Kim, T.K., Kittler, J., Cipolla, R.: Discriminative learning and recognition of image set classes using canonical correlations. Pattern Analysis and Machine Intelligence, IEEE Transactions on **29** (2007) 1005–1018
22. Chang, C.C., Lin, C.J.: Libsvm: a library for support vector machines. ACM Transactions on Intelligent Systems and Technology (TIST) **2** (2011) 27
23. Hotelling, H.: Relations between two sets of variates. Biometrika **28** (1936) 321–377
24. Witten, D.M., Tibshirani, R., et al.: Extensions of sparse canonical correlation analysis with applications to genomic data. Statistical applications in genetics and molecular biology **8** (2009) 1–27
25. Rupnik, J., Shawe-Taylor, J.: Multi-view canonical correlation analysis. In: Conference on Data Mining and Data Warehouses (SiKDD 2010). (2010) 1–4
26. Loog, M., van Ginneken, B., Duin, R.P.: Dimensionality reduction of image features using the canonical contextual correlation projection. Pattern Recognition **38** (2005) 2409–2418
27. Silberman, N., Fergus, R.: Indoor scene segmentation using a structured light sensor. In: Computer Vision Workshops (ICCV Workshops), 2011 IEEE International Conference on, IEEE (2011) 601–608
28. Lai, K., Bo, L., Ren, X., Fox, D.: A large-scale hierarchical multi-view rgb-d object dataset. In: Robotics and Automation (ICRA), 2011 IEEE International Conference on, IEEE (2011) 1817–1824
29. Brown, M., Susstrunk, S.: Multi-spectral sift for scene category recognition. In: Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on, IEEE (2011) 177–184
30. Oliva, A., Torralba, A.: Modeling the shape of the scene: A holistic representation of the spatial envelope. International journal of computer vision **42** (2001) 145–175
31. Lazebnik, S., Schmid, C., Ponce, J.: Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In: Computer Vision and Pattern Recognition, 2006 IEEE Computer Society Conference on. Volume 2., IEEE (2006) 2169–2178
32. Bo, L., Ren, X., Fox, D.: Unsupervised feature learning for rgb-d based object recognition. ISER, June (2012)