

# Learning to Recognize Objects from Unseen Modalities

C. Mario Christoudias<sup>1</sup>, Raquel Urtasun<sup>2</sup>, Mathieu Salzmann<sup>1</sup> and Trevor Darrell<sup>1</sup>

<sup>1</sup>UC Berkeley EECS & ICSI

<sup>2</sup>TTI Chicago







**Abstract.** In this paper we investigate the problem of exploiting multiple sources of information for object recognition tasks when additional modalities that are not present in the labeled training set are available for inference. This scenario is common to many robotics sensing applications and is in contrast with the assumption made by existing approaches that require at least some labeled examples for each modality. To leverage the previously unseen features, we make use of the unlabeled data to learn a mapping from the existing modalities to the new ones. This allows us to predict the missing data for the labeled examples and exploit all modalities using multiple kernel learning. We demonstrate the effectiveness of our approach on several multi-modal tasks including object recognition from multi-resolution imagery, grayscale and color images, as well as images and text. Our approach outperforms multiple kernel learning on the original modalities, as well as nearest-neighbor and bootstrapping schemes.

## 1 Introduction

Recent advances in object recognition have shown that exploiting multiple sources of information could significantly improve recognition performance. This is the case when relying either on different image features [1–3], or on multiple modalities such as images and text [4–6]. Typically, the different inputs are combined either via kernels (i.e., multiple kernel learning) [1–3, 7] or by voting schemes [8, 9]. While these techniques have proven successful, they assume that all the modalities (features) are present during both training and inference.

However, this assumption is often violated in more dynamic scenarios where new modalities are added during inference. This is common in robotics applications, where a robot can be equipped with new sensors (e.g., high resolution cameras, laser range finders). Even though existing techniques can handle a certain degree of missing data, they all require some labeled examples for each modality. As a consequence, the only way for them to exploit these new modalities is to manually label some examples and re-train the classifier.

In this paper, we tackle the problem of exploiting novel modalities that are present only at test time for which no labeled samples are provided (see Fig. 1). This scenario is particularly challenging since the number of unlabeled examples might be small. To be able to leverage the previously unseen features, we

	Conventional			Our approach			
Labeled training set							Low resolution
Unlabeled test set							Low resolution
							High resolution
Performance	75%	5%	33%	92%	29%	58%	

**Fig. 1. Exploiting unseen modalities:** In this paper we propose a new object recognition approach that can leverage additional modalities that are fully unlabeled. The example above illustrates how additional unlabeled high-resolution images let us significantly boost the classification performance over the low-resolution feature channel. Similar behavior is shown in our experiments when adding unlabeled color images to grayscale ones, and when using text in conjunction with images.

assume that the conditional distribution of the new modalities given the existing ones is stationary. This is similar in spirit to the assumption typically made by semi-supervised learning techniques that the labeled and unlabeled examples are drawn from the same distribution. This lets us exploit the unlabeled data to learn a non-linear mapping from the existing modalities to the novel ones. From the resulting mapping, we “hallucinate” the missing data on the labeled training examples. This allows us to exploit the full potential of multiple kernel learning by using both old and new modalities.

As a result, our classifier improves over the original one by effectively making use of all the available modalities while avoiding the burden of manually labeling new examples. This is of crucial importance to make recognition systems practical in applications such as personal robotics, where we expect the users to update their robot, but cannot expect them to label a set of examples each time a new sensor is added.

We demonstrate the effectiveness our approach on a variety of real-world tasks: we exploit unlabeled high-resolution images to improve webcam object recognition, we utilize unlabeled color images for grayscale object recognition, we use unlabeled text to improve visual classification, and we exploit unlabeled images for sense disambiguation in the text domain. In all these scenarios we show that our method significantly outperforms multiple kernel learning on the labeled modalities, as well as nearest-neighbor and bootstrapping schemes.

## 2 Related work

Many techniques have been proposed that exploit multiple feature cues or information sources for performing object recognition. A popular approach is to use multiple kernel learning (MKL) either in an SVM framework [1, 3] or in a

Gaussian processes probabilistic framework [2]. Voting schemes have also been proposed for multi-feature object recognition [9, 8]. In [9] the implicit shape model (ISM) was extended to include multiple features, while in [8] a naive-Bayes nearest-neighbor classifier within a voting-based scheme was utilized. These multiple feature recognition approaches have been shown to be highly beneficial and lead to state-of-the-art performance on several challenging problems [7]. However, these approaches have focused on supervised or semi-supervised scenarios where at least some labels are provided for each modality, and cannot exploit additional unsupervised modalities available only at test time.

Semi-supervised multi-view learning approaches have been used to exploit both labeled and unlabeled data. In [10] co-training was used to learn an object detector. Bayesian co-training was explored in [11] for instance-level object recognition. Similarly, multi-view bootstrapping schemes were used in [12] to transcribed speech and video features for video concept detection, and in [13] to learn audio-visual speech and gesture classifiers. Still, most of these approaches make the assumption that at least some labels are provided for each modality. The exception being cross-modal bootstrapping [13] that can leverage a classifier from a single view to learn a multi-view classifier. However, as demonstrated in our experiments this approach does not take full advantage of the unlabeled modalities.

Methods for learning a joint latent space from multiple modalities have also been proposed. In [4, 5] latent Dirichlet allocation (LDA) was used to perform visual sense disambiguation using unsupervised text and images. In [14] a transfer learning approach was proposed to learn a discriminatively trained latent space over images and their captions. Such methods can be seen as complementary to our approach in that we also exploit a form of information transfer between modalities to infer the missing ones. Yet, to our knowledge, no previous approach has considered the problem of having modalities for which no labeled examples are provided, and this is the first attempt to do so.

The most related work to ours is probably [6] where they employ a nearest-neighbor approach to infer text histograms from images using a large external collection of images and text captured from the web. This is different from our approach in that their method is specifically designed to infer text and assumes that a very large dataset (i.e., hundreds of thousands of examples) is available. As evidenced by our experiments, our approach significantly improves over nearest-neighbor inference across a wide range of problems.

### 3 Exploiting unseen modalities

In this section, we present our approach to exploiting new modalities at test time even though there is no labeled data for them. Towards this end, we show how to hallucinate the missing modalities for the labeled examples by learning a mapping from the old modalities to the new ones from the unlabeled data. Given these hallucinated modalities, we propose a framework that combines the different sources of information using probabilistic multiple kernel learning. We then introduce a representation of the novel modalities that lets us exploit the full

potential of non-linear kernels recently developed for object recognition, while still making regression possible. Finally, we present a bootstrapping algorithm that further improves the performance of our classifier.

### 3.1 “Hallucinating” the missing modalities

To leverage the availability of fully unsupervised modalities for classification, we propose to infer these missing modalities for the labeled examples and use them in conjunction with the old modalities in a probabilistic multiple kernel learning framework. In this section we show how to hallucinate the missing modalities.

More formally, let  $\mathbf{X} = [\mathbf{X}^{(1)}, \dots, \mathbf{X}^{(D)}]$  be the set of training inputs for the  $D$  modalities present in both the labeled and unlabeled datasets, with  $\mathbf{X}^{(i)} = [\mathbf{x}_1^{(i)}, \dots, \mathbf{x}_{N_{train}}^{(i)}]^T$  the  $N_{train}$  training points for the  $i$ -th modality, and let  $\bar{\mathbf{X}} = [\bar{\mathbf{X}}^{(1)}, \dots, \bar{\mathbf{X}}^{(D)}]$  be the test examples, with  $\bar{\mathbf{X}}^{(i)} = [\bar{\mathbf{x}}_1^{(i)}, \dots, \bar{\mathbf{x}}_{N_{test}}^{(i)}]^T$ . Let  $\bar{\mathbf{Z}} = [\bar{\mathbf{Z}}^{(1)}, \dots, \bar{\mathbf{Z}}^{(M)}]$  be the  $M$  new modalities present only in the unlabeled test set, such that  $\bar{\mathbf{Z}}^{(i)} = [\bar{\mathbf{z}}_1^{(i)}, \dots, \bar{\mathbf{z}}_{N_{test}}^{(i)}]^T$ .

In order to exploit the new modalities that are only present at test time and for which we do not have any labeled examples, we assume that the conditional distribution of the new modalities given the labeled ones is stationary, i.e., the same at training and inference. This is similar in spirit to the standard assumptions of semi-supervised learning techniques, and lets us rely on the concept of mean imputation typically used when dealing with missing data. However, here we assume that only a small amount of unlabeled data is available to learn the mapping from the known modalities  $\mathbf{x}$  to the new modalities  $\mathbf{z}$ . This makes the problem more challenging, since simple methods such as nearest-neighbors (NN) require large collections of examples for accurate prediction.

To overcome this issue, we rely on Gaussian processes (GPs) which have proven effective when trained from a small number of examples [15]. This is due to the fact that they marginalize among all possible non-linear mappings defined by the kernel function. In particular, we utilize a GP to learn the mapping from the known modalities to the missing ones. Note that unlike for the classification task, when hallucinating the new modalities the unlabeled examples are used as training data, since for those both  $\bar{\mathbf{Z}}$  and  $\bar{\mathbf{X}}$  are known. Under this model, the likelihood can be expressed as

$$p(\bar{\mathbf{Z}}|\bar{\mathbf{X}}) = \prod_{m=1}^M \prod_{i=1}^{S_m} p(\bar{\mathbf{Z}}_{:,i}^{(m)}|\bar{\mathbf{X}}) = \prod_{m=1}^M \prod_{i=1}^{S_m} \mathcal{N}(\bar{\mathbf{Z}}_{:,i}^{(m)}; 0, \mathbf{K}^x), \quad (1)$$

where  $S_m$  is the dimensionality of the  $m$ -th new modality. The elements of the kernel associated with the labeled modalities  $\mathbf{K}^x$  are computed by kernel combination as

$$\mathbf{K}_{i,j}^x = \sum_{m=1}^D \alpha_m k^x(\bar{\mathbf{x}}_i^{(m)}, \bar{\mathbf{x}}_j^{(m)}), \quad (2)$$

where  $\alpha_m$  are hyper-parameters of the model. In order to capture the correlations between the different output dimensions and modalities, we share kernel hyper-parameters across the different predictors.

Given the known modalities  $\mathbf{x}_i$  for a labeled example, the predictive distribution under the Gaussian process is also Gaussian and can be computed in closed form, i.e.,  $p(\mathbf{z}_i|\mathbf{x}_i, \bar{\mathbf{X}}, \bar{\mathbf{Z}}) = \mathcal{N}(\mu(\mathbf{x}_i), \sigma(\mathbf{x}_i))$ , with mean and variance

$$\mu(\mathbf{x}_i) = \mathbf{k}_i^x (\mathbf{K}^x)^{-1} \bar{\mathbf{Z}} \quad (3)$$

$$\sigma(\mathbf{x}_i) = k^x(\mathbf{x}_i, \mathbf{x}_i) - \mathbf{k}_i^x (\mathbf{K}^x)^{-1} \mathbf{k}_i^{xT}, \quad (4)$$

where  $\mathbf{k}_i^x$  is the vector obtained by evaluating the kernel function of Eq. (2) between  $\bar{\mathbf{X}}$  and  $\mathbf{x}_i$ . For each labeled example, we hallucinate the missing modalities by taking them as the mean prediction of the learned GP. The resulting hallucinated modalities predicted by the GP can then be used in conjunction with the labeled ones in the probabilistic multiple kernel learning framework described in the following section. Note that here we have made the assumption that the mapping between the old and new modalities is unimodal, i.e., can be modeled with a GP. As suggested by our results, this assumption is reasonable for a wide range of problems. In more challenging scenarios, it can easily be relaxed by using a mixture of local predictors [16].

### 3.2 Probabilistic multiple kernel learning

To exploit all available sources of information for classification, we combine the hallucinated modalities and the old ones within a probabilistic multiple kernel learning framework. In particular, we employ GPs to learn the mapping from  $(\mathbf{x}, \mathbf{z})$  to the labels  $y$ . This has been shown to perform similarly to SVM-based MKL approaches while being computationally more efficient [2]. This yields the likelihood  $p(\mathbf{y}|\mathbf{X}, \mathbf{Z}) = \mathcal{N}(\mathbf{y}; 0, \mathbf{K})$ , where  $\mathbf{y} = [y_1, \dots, y_N]^T$  are the labels for the training examples and the elements of  $\mathbf{K}$  are computed as

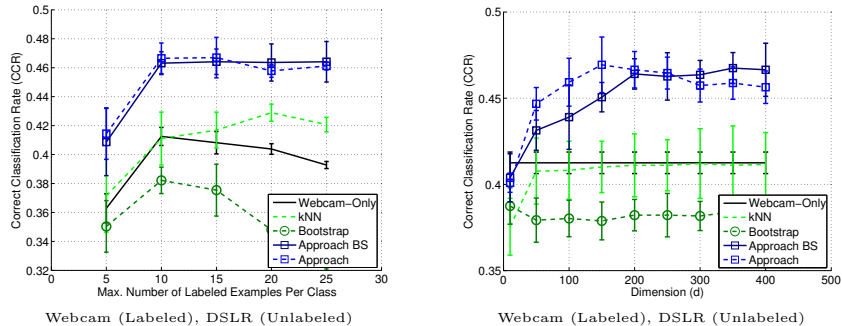
$$\mathbf{K}_{i,j} = \mathbf{K}_{i,j}^x + \sum_{m=1}^M \beta_m k^z(\mathbf{z}_i^{(m)}, \mathbf{z}_j^{(m)}), \quad (5)$$

with  $k^z(\cdot, \cdot)$  the kernel function for the unsupervised modalities.

Given new input observations  $(\bar{\mathbf{x}}, \bar{\mathbf{z}})$ , we use the mean prediction of the GP to assign a class label to the example. For multi-class problems we used a one-vs-all strategy that selects the class having the largest positive mean prediction. Note that we use a Gaussian noise model which has been shown to perform similarly to more complex noise models (e.g., probit, logit) [17].

### 3.3 A general representation of the novel modalities

While for some representations of  $\mathbf{z}$  (e.g., histograms) the above framework could be effective, we would like our MKL algorithm to be able to exploit complex non-linear kernels (e.g., Pyramid Match Kernel (PMK) [18], Spatial Pyramid [19]) that have proven successful for object recognition tasks. Note that, with these kernels,  $\mathbf{z}$  would correspond to the high-dimensional feature map which cannot be computed in practice.



**Fig. 2. Robotics dataset:** We consider the scenario where only low-resolution webcam images are labeled and an additional unlabeled high-resolution modality is available at test time using the dataset of [21]. (left) Comparison of our approach against several baselines as a function of the number of labeled examples, Approach BS is our approach with bootstrapping. (right) Accuracy as a function of K-PCA dimensionality for  $Q = 10$ . Note in both cases our approach outperforms the baselines. Error bars indicate  $\pm 1$  std.

We overcome this difficulty by learning a representation of the unlabeled modalities that is able to exploit complex non-linear kernels. To this end, we rely on Kernel PCA [20], which computes a low-dimensional representation of the possibly infinite dimensional input space by performing the eigendecomposition of the (centered) kernel matrix evaluated on the unlabeled data. This representation is interesting since it is low-dimensional and allows us to use any Mercer kernel to represent the new modalities. Our approach then proceeds as follows: First, we compute K-PCA on the new modalities and retain the first  $d$  dimensions. We then regress from the old modalities to the K-PCA representation of the new ones to hallucinate the missing data. Given the hallucinated data, we perform probabilistic multiple kernel learning with all the modalities, using a kernel computed in the K-PCA embedding for the new modalities.

### 3.4 Bootstrapping

To make further use of the unlabeled examples, we propose to use a bootstrapping strategy. At first, our multiple kernel classifier is trained on all the modalities using the hallucinated data. We then evaluate it on the unlabeled data and add the  $B$  most confident predictions per class to the set of labeled examples. For the confidence measure, we rely on the distance from the mean prediction to the predicted class label. This is similar to the concept of margin in SVMs. Note that other criteria used for active learning could be employed, e.g., uncertainty [2]. Given the old and new labeled examples, we train a new classifier and repeat the process  $T$  times.

## 4 Experimental evaluation

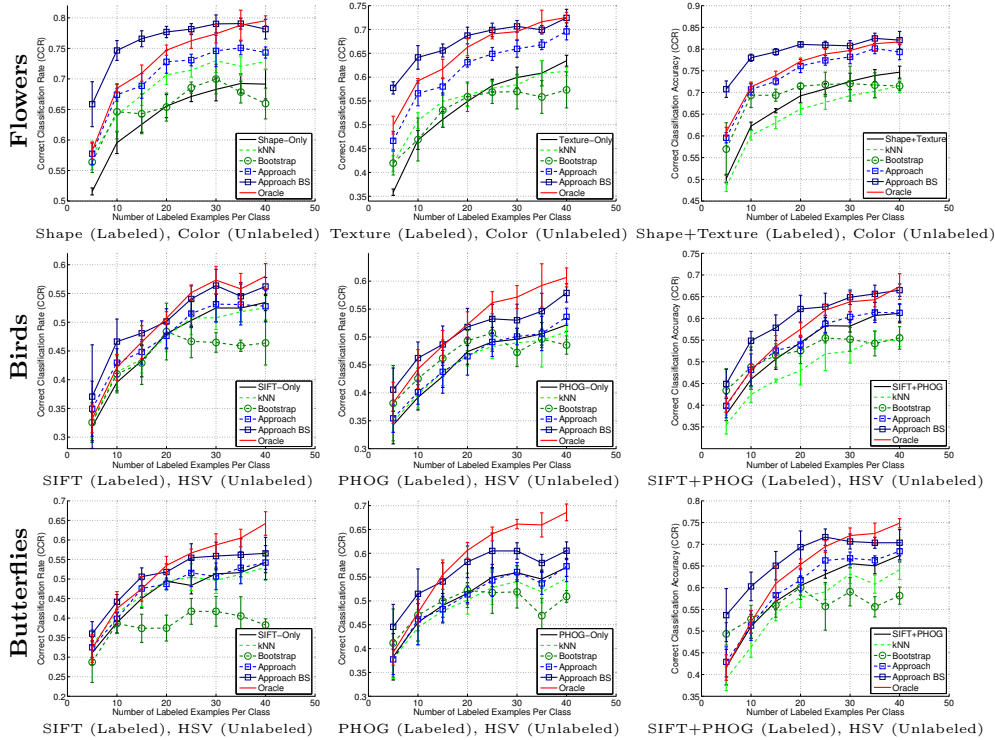
We evaluated our approach on a broad range of object recognition domains where modalities that are available at test time may not be present in the labeled training set. In particular, we first show how we can classify high-resolution images



**Fig. 3. Most confident images in the robotics application:** For each class, the top row shows the most confident images returned by the low-resolution only classifier, and the bottom row depicts similar images for our classifier trained with additional unlabeled high-resolution images. Our approach significantly improves the results for some of the classes and has less impact on others. Nonetheless, even for these classes it reduces the ambiguity, e.g., to only two different labels.

when only low-resolution webcam images are labeled. We then show how exploiting unlabeled color images in conjunction with intensity-only labeled data improves classification performance. Finally, we address the problem of classification of text and image datasets, where only either text or images are labeled. Across all of these real-world problem domains, our approach achieves a significant performance boost by exploiting the additional test modalities without requiring any new labeled samples. The paper code and databases are available online at <http://www.eecs.berkeley.edu/Research/Projects/CS/vision/shape/msorec/>.

**Baselines:** On each dataset we compare against two baselines. The first one employs  $k$  nearest-neighbor ( $k$ -NN) to infer the missing modalities by averaging across the  $k$  neighbors. This approach is akin to [6] where they rely on  $k$ -NN to infer text features from a large collection of web images. On each dataset

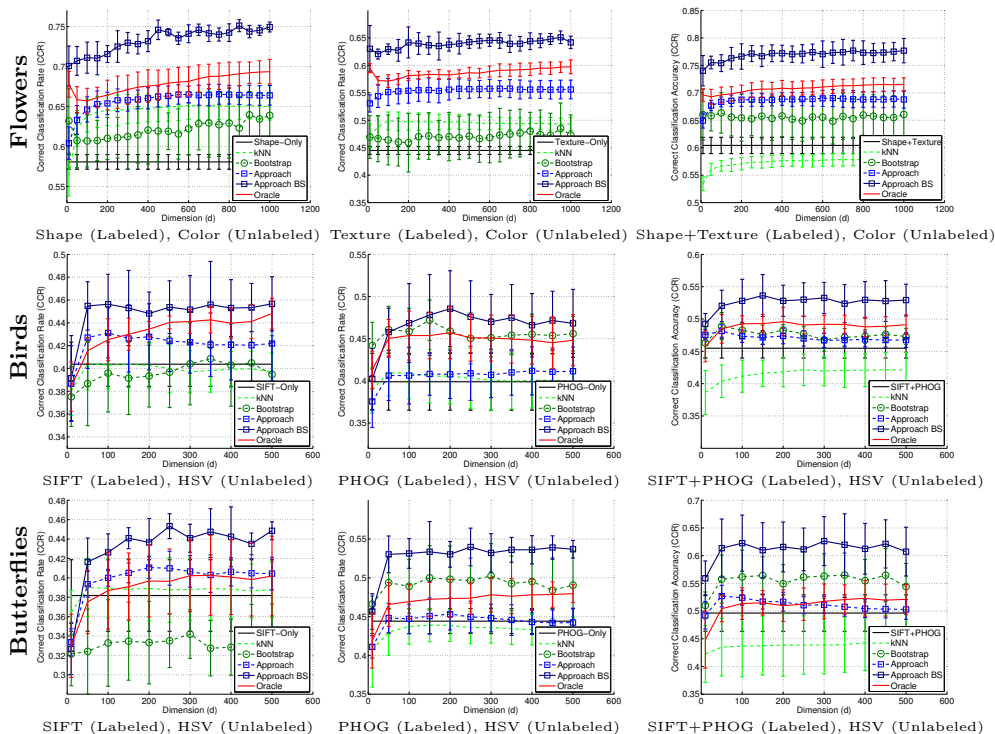


**Fig. 4. Using unlabeled color images for grayscale object recognition:** For each dataset, we show (left,middle) the performance of our approach using only a single intensity feature, and (right) using both of the available intensity features. Note that our approach achieves a significant performance boost over intensity-only performance and outperforms the other baselines. Performance is shown across different training sizes with  $d = 200$ . The error bars indicate  $\pm 1$  std.

we chose the  $k$  that gave the best performance. Note that the baseline makes use of the test data to select the best parameter  $k$ , while our approach does not. The second baseline (Bootstrap) exploits the additional test modalities by cross-modal bootstrapping analogous to [13]; an initial seed set is formed by labeling  $B_{init}$  examples per class using the single-view classifier, and the same bootstrapping strategy used by our approach is then applied. We also show single-view and oracle performance for each dataset, where the oracle makes use of the ground-truth features on the missing modalities for the labeled training set.

**Experimental setup:** For the bootstrapping baseline and our approach, we used  $B = 2$ , and  $T = 10$  across all datasets, and  $B_{init} = 10$  for the baseline. We used RBF kernels whose bandwidths were set to the mean squared distance over the dataset to compute K-PCA and for GP regression. Additionally, we set the Gaussian noise variance to 0.01. For all but the Mouse dataset, we used  $d = 200$ . For this dataset we used  $d = 20$ , since the number of examples is much smaller than in the other datasets. In the case of multiple kernel classifiers, we set  $\alpha_m$



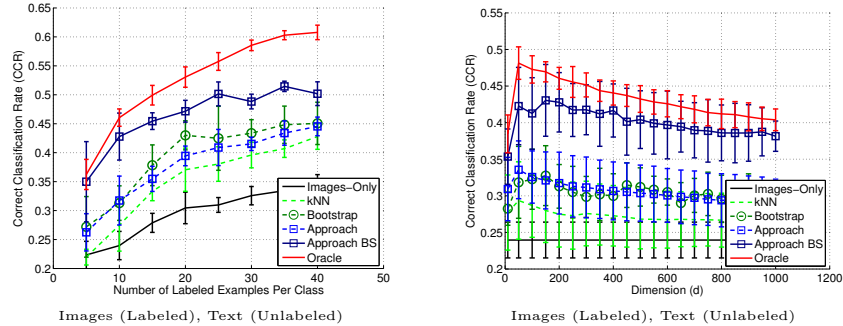


**Fig. 5. Using unlabeled color images for grayscale object recognition:** Classification error as a function of the K-PCA dimensionality for  $Q = 10$ . The plots show the performance using a single labeled intensity feature (left,middle) and using both available intensity features (right) to classify the color images. Our approach is insensitive to the choice of dimensionality and improves over the baselines.

and  $\beta_m$  to be  $1/V$  where  $V$  is the total number of views. We did not optimize  $\alpha_m$  and  $\beta_m$ , since it has been shown that averaging the kernels yields similar performance [2, 22]. We ran each experiment on 5 random splits of the data into training and test sets where for each split we used  $Q$  labeled examples per class. The performance of each approach was evaluated using the Correct Classification Rate (CCR) defined as the total number of correctly classified examples divided by the total number of examples.

#### 4.1 Multi-sensor object recognition

We first consider a multiple sensor robotics scenario and show that our approach lets us leverage the better quality of high resolution images even though only low resolution images were labeled. We used a subset of the images from the dataset of [21] yielding 783 low resolution and 486 high resolution images of 30 office object categories captured using a webcam and a DSLR camera. We split the dataset into webcam-only images and webcam+DSLR image pairs that depict an object from similar viewpoints. This resulted in a total of 368 webcam images used as labeled examples and 415 webcam+DSLR image pairs that we treated



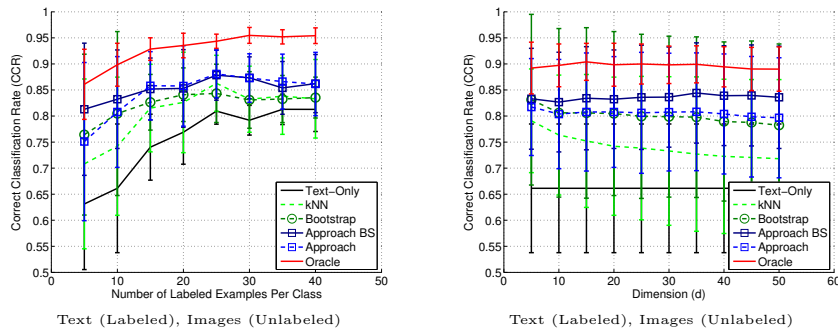
**Fig. 6. Using unlabeled text to improve visual classification:** We focus on a recognition task that exploits an additional text modality that is only present at test time (no labeled data is available) to improve the performance of a visual classifier. We used the dataset of [5] that consists of a set of noisily labeled images obtained from an online keyword search. Using unlabeled text features our approach is able to significantly improve performance over the weakly supervised image classifier. Performance is shown across different training set sizes with  $d = 200$  (left) and across different K-PCA dimensions with  $Q = 10$  (right). Error bars indicate  $\pm 1$  std.

as our unlabeled set. For each sensor type, we used PHOG features [23] with 4 pyramid levels and 8 histogram bins per cell over 360 degrees. We then applied PCA to those vectors and retained 95% of the variance to form the final feature vector.

This dataset is fairly challenging as it consists of a wide variety of object categories and appearances taken from a sparse set of varying viewpoints. Fig. 2 (left) depicts performance as a function of the training set size. For this dataset, at most  $Q$  labeled examples were retained per class as some classes had fewer than  $Q$  webcam images across the different training set sizes. Since the number of unlabeled examples is fairly small in this dataset, the performance of the  $k$ -NN baseline is poor. For similar reasons, the cross-modality bootstrap baseline is unable to improve over the weak performance of the webcam-only classifier. In contrast, our approach that infers the missing DSLR modality on the labeled training set results in a stronger multi-view classifier. Fig. 2 (right) depicts performance with varying K-PCA dimensionality for  $Q = 10$  training samples per class. Our approach proves to be fairly insensitive to the choice of  $d$  and outperforms the baselines across a wide range of dimensionalities of the embedding. In Fig. 3, we compare the most confident images returned by the webcam-only classifier and by our approach. As can be observed from the images, our classifier is able to avoid some of the mistakes of the webcam-only baseline.

#### 4.2 Using unlabeled color images for grayscale object recognition

We now illustrate how our approach is capable of exploiting the additional information contained in unlabeled color images to improve the performance of a grayscale object classifier, when only grayscale examples are labeled. This is a plausible scenario in robotics applications where robots are equipped with high-performance (e.g., hyperspectral) cameras. For this task we used three datasets

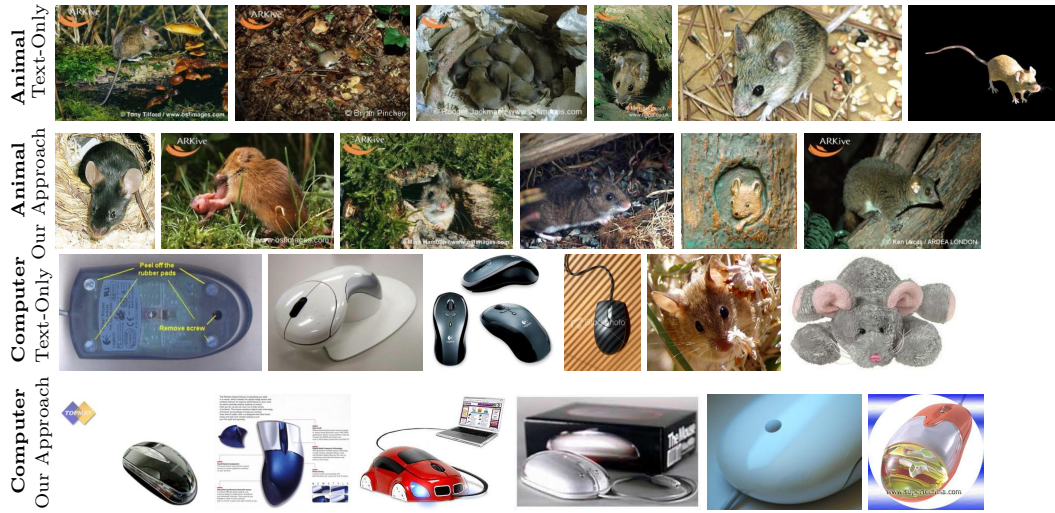


**Fig. 7. Using unlabeled images for sense disambiguation in text classification:** We tackle the problem of sense disambiguation from a labeled set of text examples and a set of unlabeled images that are available only at test time. We used the dataset of [4] to disambiguate the two meanings of the word mouse that pertain to the animal and the computer device classes. Our approach is able to significantly improve over a text-only classifier,  $k$ -NN and bootstrap baselines using only unlabeled images. Performance is shown across different training set sizes with  $d = 20$  (right) and across different dimensions with  $Q = 10$  (left). Error bars indicate  $\pm 1$  std.

of natural object categories, where color features are relevant for classification. The first dataset is the Oxford Flowers dataset [3] that is comprised of 80 images of 17 flower categories. For this dataset, the authors have provided  $\chi^2$  distance matrices over three visual feature channels: Color, Shape and Texture. The other two datasets are comprised of butterfly and bird image categories respectively [24, 25]. The Birds dataset contains 6 classes with 100 images per class and the Butterflies dataset has 619 images of 7 classes. For these datasets we extracted dense SIFT [26], PHOG [23] and HSV color features [3]. Images were compared using  $\chi^2$  distances for SIFT and HSV features and  $L_2$  distance for PHOG.

Fig. 4 displays performance on the three datasets as a function of the training set size. Note that unlike  $k$ -NN, our approach is able to obtain a good estimate of the missing color modality from few training samples, and significantly improves over grayscale-only performance. The conventional bootstrapping baseline is also unable to achieve a significant improvement and often underperforms grayscale-only performance. In contrast, an even greater improvement is achieved with our approach when used in combination with bootstrap (Approach BS) that often matches or even improves upon the supervised oracle. Fig. 5 depicts performance for the three datasets as a function of the dimensionality of the K-PCA embedding for  $Q = 10$ . Note that the performance of our approach is relatively insensitive to the choice of dimensionality.

For completeness of our evaluation, we compared our approach and the baselines over all the possible feature combinations, even though these combinations do not necessarily represent real-world scenarios, e.g., having HSV and PHOG at test time but only labeled SIFT. Similar performance as to that of the above cases was observed. These plots are available at the project webpage listed above.



**Fig. 8. Most confident images in sense disambiguation:** As in Fig. 3, the top row of each class shows the most confident images returned by the classifier built from labeled text features, and the bottom row depicts the result of our classifier when using images as an additional unsupervised modality. Note that, for the animal meaning, both approaches perform similarly whereas our classifier outperforms the text-only one on the computer sense.

### 4.3 Using unlabeled text to improve visual classification

Next, we focused on the task of leveraging unlabeled text features to improve over image-only object categorization. Our labeled set consists of a set of images collected using keyword search from an image search engine. Such a dataset can be considered as weakly supervised in that many images returned for a given object keyword can be only very loosely related to the target category. In the test set each image is accompanied with text (e.g., extracted from the webpage). We used the Office dataset of [5] that consists of text and images for 10 object categories. As this dataset is fairly large, we considered a subset of the data and randomly chose 200 examples per class to form our evaluation set. We used the same features as [5], which consists of histograms of SIFT features and word histograms. The histograms were compared using  $\chi^2$  distances.

Fig. 6 depicts performance as a function of the training set size for  $d = 200$ , and as a function of the dimensionality of the K-PCA embedding for  $Q = 10$ . As expected, due to the large amount of noise, performance using images alone is fairly weak. In contrast, our approach is able to leverage the additional unlabeled text modality and significantly improves recognition performance over the weakly supervised image classifier. When combined with bootstrapping, it nearly matches oracle performance. Both the  $k$ -NN and the cross-modality bootstrap baselines are also able to improve over image-only performance, although they do not perform as well as our method. Similarly as before, we can see that our approach is rather insensitive to the choice of  $d$ .

#### 4.4 Using unlabeled images for sense disambiguation in text

We now consider the problem of exploiting unlabeled images to disambiguate the sense of classes described by labeled text features only. Sense disambiguation is important when each object category can pertain to multiple visual senses [4, 5]. For example, the keyword MOUSE can pertain to the animal or the computer device. Our goal is to use unsupervised images to improve the performance of a text-only classifier to discriminate polysemous object categories. We used a subset of the dataset of [4] that consists of about 100 examples per class, selected to contain images only of the target senses. Each image is represented using a histogram of dense SIFT features and each text document is summarized into a word histogram. Histograms are compared using the  $\chi^2$  distance.

Fig. 7 depicts performance as a function of the training set size for  $d = 200$  and of the dimensionality of the embedding for  $Q = 10$ . Although performance with text-only features is fairly good, the addition of the unlabeled visual modality significantly improves performance. Once again our approach outperforms the  $k$ -NN and cross-modality bootstrap baselines. Fig. 8 depicts some of the most confident images obtained for each class by either the text-only classifier, or by our approach. Note that our approach is able to avoid some of the mistakes made by the text-only classifier.

## 5 Conclusion

In this paper we have investigated the problem of exploiting multiple sources of information when some of the modalities do not have any labeled examples, i.e., are only present at test time. Assuming that the conditional distribution of novel modalities given old ones is stationary, we have shown how to make use of the unlabeled data to learn a non-linear mapping that hallucinates the new modalities for the labeled examples. Furthermore, we have shown how to learn low-dimensional representations that allow us to exploit complex non-linear kernels developed for object recognition. Finally, our approach is able to employ multiple kernel learning with all the modalities as well as a bootstrapping strategy that further improved performance. We have demonstrated the effectiveness of our approach on several tasks including object recognition from intensity and color cues, text and images, and multi-resolution imagery. In the future we plan to investigate complementary techniques for inferring the missing views that include learning a shared latent space and the use of local GPs to cope with multi-modal output spaces.

## References

1. Varma, M., Ray, D.: Learning the discriminative power-invariance trade-off. In: ICCV. (2007)
2. Kapoor, A., Graumann, K., Urtasun, R., Darrell, T.: Gaussian processes for object categorization. IJCV (2009)
3. Nilsback, M.E., Zisserman, A.: A visual vocabulary for flower classification. In: CVPR. (2006)

4. Saenko, K., Darrell, T.: Unsupervised learning of visual sense models for polysemous words. In: NIPS. (2008)
5. Saenko, K., Darrell, T.: Filtering abstract senses from image search results. In: NIPS. (2009)
6. Wang, G., Hoiem, D., Forsyth, D.: Building text features for object image classification. In: CVPR. (2009)
7. Vedaldi, A., Gulshan, V., Varma, M., Zisserman, A.: Multiple kernels for object detection. In: CVPR. (2009)
8. Boiman, O., Shechtman, E., Irani, M.: In defense of nearest-neighbor based image classification. In: CVPR. (2008)
9. Leibe, B., Mikolajczyk, K., Schiele, B.: Segmentation based multi-cue integration for object detection. In: BMVC. (2006)
10. Levin, A., Viola, P., Freund, Y.: Unsupervised improvement of visual detectors using co-training. In: ICCV. (2003)
11. Christoudias, C.M., Urtasun, R., Kapoor, A., Darrell, T.: Co-training with noisy perceptual observations. In: CVPR. (2009)
12. Yan, R., Naphade, M.: Semi-supervised cross feature learning for semantic concept detection in videos. In: CVPR. (2005)
13. Christoudias, C.M., Saenko, K., Morency, L.P., Darrell, T.: Co-adaptation of audio-visual speech and gesture classifiers. In: ICMI. (2006)
14. Quattoni, A., Collins, M., Darrell, T.: Learning visual representations using images with captions. In: CVPR. (2007)
15. Urtasun, R., Fleet, D., Hertzmann, A., Fua, P.: Priors for people tracking from small training sets. In: ICCV. (2005)
16. Urtasun, R., Darrell, T.: Sparse probabilistic regression for activity-independent human pose inference. In: CVPR. (2008)
17. Urtasun, R., Quattoni, A., Lawrence, N., Darrell, T.: Transferring nonlinear representations using gaussian processes with a shared latent space. Technical report, MIT (2008)
18. Grauman, K., Darrell, T.: The pyramid match kernel: Discriminative classification with sets of image features. In: ICCV. (2005)
19. Lazebnik, S., Schmid, C., Ponce, J.: Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In: CVPR. (2006)
20. Scholkopf, B., Smola, A., Muller, K.: Kernel principal component analysis. In: ICANN. (1997)
21. Saenko, K., Kulis, B., Fritz, M., Darrell, T.: Adapting visual category models to new domains. In: ECCV. (2010)
22. Gehlert, P., Nowozin, S.: Learning image similarity from flickr groups using stochastic intersection kernel machines. In: ICCV. (2009)
23. Bosch, A., Zisserman, A., Munoz, X.: Image classification using random forest and ferns. In: ICCV. (2007)
24. Lazebnik, S., Schmid, C., Ponce, J.: Semi-local affine parts for object recognition. In: BMVC. (2004)
25. Lazebnik, S., Schmid, C., Ponce, J.: A maximum entropy framework for part-based texture and object recognition. In: CVPR. (2005)
26. Lowe, D.G.: Distinctive image features from scale-invariant keypoints. IJCV (2004)