From Virtual to Reality: Fast Adaptation of Virtual Object Detectors to Real Domains

Baochen Sun
http://www.cs.uml.edu/~bsun
Kate Saenko
http://www.cs.uml.edu/~saenko

Computer Science Department
University of Massachusetts Lowell
Lowell, Massachusetts, US

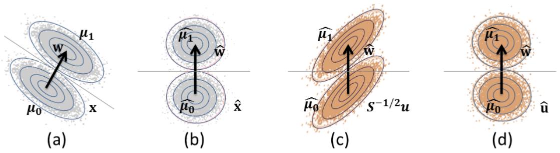

Figure 1: (a) Applying a linear classifier $\mathbf{w}$ learned by LDA to source data $\mathbf{x}$ is equivalent to (b) applying classifier $\hat{\mathbf{w}} = \mathbf{S}^{-1/2}\mathbf{w}$ to decorrelated points $\mathbf{S}^{-1/2}\mathbf{x}$. (c) However, target points $\mathbf{u}$ may still be correlated after $\mathbf{S}^{-1/2}\mathbf{u}$, hurting performance. (d) Our method uses target-specific covariance to obtain properly decorrelated $\hat{\mathbf{u}}$.

Abstract. The most successful 2D object detection methods require a large number of images annotated with object bounding boxes to be collected for training. We present an alternative approach that trains on virtual data rendered from 3D models, avoiding the need for manual labeling. Growing demand for virtual reality applications is quickly bringing about an abundance of available 3D models for a large variety of object categories. While mainstream use of 3D models in vision has focused on predicting the 3D pose of objects, we investigate the use of such freely available 3D models for multicategory 2D object detection. To address the issue of dataset bias that arises from training on virtual data and testing on real images, we propose a simple and fast adaptation approach based on decorrelated features.

Background. In recent years, use of the linear SVM with Histogram of Gradients (HOG) as the features has emerged as the predominant object detection paradigm. Yet, as observed by Hariharan *et al.* [3], training SVMs can be expensive, especially because it usually involves costly rounds of hard negative mining. Furthermore, the training must be repeated for each object category, which makes it scale poorly with the number of categories. Hariharan et al. proposed a much more efficient alternative using Linear Discriminant Analysis (LDA). LDA is a well-known linear classifier that models the training set of examples $\mathbf{x}$ with labels $y \in \{0,1\}$ as being generated by $p(\mathbf{x},y) = p(\mathbf{x}|y)p(y)$. $p(y)$ is the prior on class labels and the class-conditional densities are normal distributions $p(\mathbf{x}|y) = N(\mathbf{x};\mu^y,\mathbf{S})$, where the feature vector covariance $\mathbf{S}$ is assumed to be the same for both positive and negative (background) classes. In our case, the feature is represented by $\mathbf{x} = \phi(I,b)$. The resulting classifier is given by The innovation in [3] was to re-use $\mathbf{S}$ and $\mu_\mathbf{0}$, the background mean, for all categories, reducing the task of learning a new category model to computing the average positive feature, $\mu_\mathbf{1}$. This was accomplished by calculating $\mathbf{S}$ and $\mu_\mathbf{0}$ for the largest possible window and subsampling to estimate all other smaller window sizes. Also, $\mathbf{S}$ was shown to have a sparse local structure, with correlation falling off sharply beyond a few nearby image locations. LDA was shown in [3] to have competitive performance to SVM, and can be implemented both as an exemplar-based [4] or as deformable parts model (DPM) [1].

Approach. We observe that estimating global statistics $\mathbf{S}$ and $\mu_\mathbf{0}$ once and re-using them for all tasks may work when training and testing in the same domain, but in our case, the virtual training data is likely to have different statistics from the target real data. Figure 2 illustrates the effect of centering and decorrelating a positive mean using global statistics from the wrong domain. The effect is clear: important discriminative information is removed while irrelevant structures are not.

Based on this observation, we propose an adaptive decorrelation approach to detection. Assume that we are given labeled training data $\{\mathbf{x},y\}$ in the source domain (*e.g.* virtual images rendered from 3D models), and unlabeled examples $\mathbf{u}$ in the target domain (*e.g.* real images collected in an office environment). Evaluating the scoring function $f_\mathbf{w}(\mathbf{x})$ in the source domain is equivalent to first decorrelating the training features $\hat{\mathbf{x}} = \mathbf{S}^{-1/2}\mathbf{x}$, computing their positive and negative class means $\hat{\mu_\mathbf{1}} = \mathbf{S}^{-1/2}\mu_\mathbf{1}$ and $\hat{\mu_\mathbf{0}} = \mathbf{S}^{-1/2}\mu_\mathbf{0}$ and then projecting the decorrelated feature onto the decorrelated difference between means, $f_\mathbf{w}(\mathbf{x}) = \hat{\mathbf{w}}^T\hat{\mathbf{x}}$, where $\hat{\mathbf{w}} = (\hat{\mu_\mathbf{1}} - \hat{\mu_\mathbf{0}})$. This is illustrated in Figure 1(a-b). However, as we saw in Figure 2, the assumption that the input is properly decorrelated does not hold if the input comes from a target domain with a different covariance structure. Figure 1(c) illustrates this case, showing that $\mathbf{S}^{-1/2}\mathbf{u}$ does not have isotropic covariance. Therefore, $\mathbf{w}$ cannot be used directly.

We may be able to compute the covariance of the target domain on the unlabeled target points $\mathbf{u}$, but not the positive class mean. Therefore, we would like to re-use the decorrelated mean difference $\hat{\mathbf{w}}$, but adapt to the covariance of the target domain. In this paper, we make the assumption that the difference between positive and negative means is the same in the source and target.

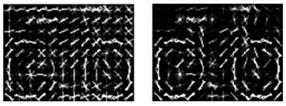

Figure 2: Mean bicycle decorrelated with mismatched-domain covariance (left) vs. with same-domain covariance (right).

Let the estimated target covariance be $\mathbf{T}$. We first decorrelate the target input feature with its inverse square root, and then apply $\hat{\mathbf{w}}$ directly, as shown in Figure 1(d). The resulting scoring function is $f_{\hat{\mathbf{w}}}(\mathbf{u}) = \left((\mathbf{T}^{-1/2})^T\mathbf{S}^{-1/2}(\mu_\mathbf{1} - \mu_\mathbf{0})\right)^T\mathbf{u}$. This corresponds to a transformation of $(\mathbf{T}^{-1/2})^T(\mathbf{S}^{-1/2})$ instead of the original whitening $\mathbf{S}^{-1}$ being applied to the difference between means to compute $\mathbf{w}$. Note that if source and target domains are the same, then $(\mathbf{T}^{-1/2})^T(\mathbf{S}^{-1/2})$ equals to $\mathbf{S}^{-1}$ since $\mathbf{S}$ is positive definite.

In practice, either the source or the target component of the above transformation may also work, or even statitstics from similar domains. However, as shown by our experiments, dissimilar domain statistics can significantly hurt performance. Furthermore, if either source or target has only images of the positive category available, and cannot be used to properly compute background statistics, the other domain can still be used.

We also extend our approach to supervised adaptation when a few labeled examples are available in the target domain. Following [2], a simple adaptation method is used whereby the template learned on source positives is combined with a template learned on target positives, using a weighted linear combination. The key difference with our approach is that the target template uses target-specific statistics. In [2], the author uses the same background statistics as [3] which were estimated on 10,000 natural images from the PASCAL VOC 2010 dataset. Based on our analysis, even though these background statistics were estimated from a very large amount of real image data, they will not work for all domains. Our results confirms this claim.

We evaluate our technique by training on virtual labeled examples and testing on real images from a benchmark domain adaptation dataset. We compare two kinds of virtual data, one rendered with real-image textures and one without. The evaluation demonstrates that with our method, performance of classifiers trained on virtual data is comparable to that of classifiers trained on large-scale real image domains.

[1] Pedro F Felzenszwalb, Ross B Girshick, David McAllester, and Deva Ramanan. Object detection with discriminatively trained part-based models. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 32(9):1627–1645, 2010.

[2] Daniel Goehring, Judy Hoffman, Erik Rodner, Kate Saenko, and Trevor Darrell. Interactive adaptation of real-time object detectors. In *International Conference on Robotics and Automation (ICRA)*, 2014.

[3] Bharath Hariharan, Jitendra Malik, and Deva Ramanan. Discriminative decorrelation for clustering and classification. In *Computer Vision–ECCV 2012*, pages 459–472. Springer, 2012.

[4] Tomasz Malisiewicz, Abhinav Gupta, and Alexei A Efros. Ensemble of exemplar-svms for object detection and beyond. In *Computer Vision (ICCV), 2011 IEEE International Conference on*, pages 89–96. IEEE, 2011.