Multi-task Bilinear Classifiers for Visual Domain Adaptation

Jiaolong Xu
Computer Vision Center, Barcelona, Spain
jiaolong@cvc.uab.es

Sebastian Ramos
Computer Vision Center, Barcelona, Spain
sramosp@cvc.uab.es

Xu Hu
EECS department, Oregon State University, USA
huxu@onid.orst.edu

David Vázquez
Computer Vision Center, Barcelona, Spain
dvazquez@cvc.uab.es

Antonio M. López
Computer Vision Center and
C. Sc. Dpt. Univ. Autònoma de Barcelona, Spain
antonio@cvc.uab.es

Abstract

We propose a method that aims to lessen the significant accuracy degradation that a discriminative classifier can suffer when it is trained in a specific domain (*source domain*) and applied in a different one (*target domain*). The principal reason for this degradation is the discrepancies in the distribution of the features that feed the classifier in different domains. Therefore, we propose a domain adaptation method that maps the features from the different domains into a common subspace and learns a discriminative domain-invariant classifier within it. Our algorithm combines bilinear classifiers and multi-task learning for domain adaptation. The bilinear classifier encodes the feature transformation and classification parameters by a matrix decomposition. In this way, specific feature transformations for multiple domains and a shared classifier are jointly learned in a multi-task learning framework. Focusing on domain adaptation for visual object detection, we apply this method to the state-of-the-art deformable part-based model for cross domain pedestrian detection. Experimental results show that our method significantly avoids the domain drift and improves the accuracy when compared to several baselines.

1 Introduction

Developing reliable object recognition and detection systems relies mostly in the fact of training accurate vision-based object classifiers. For such training process, there has been a lot of effort in looking for good features and appropriate learning machines. In this context, most of the methods proposed for learning classifiers implicitly assume that the training and testing data are statistically similar. Unfortunately, the accuracy of such classifiers can drop significantly when the training data (*source domain*) and the application scenario (*target domain*) have inherent differences. For instance, when the training and testing data are collected using different cameras or when the object poses and views distributions are different in each dataset. To avoid this problem, it is required to adapt the classifier trained on the source domain to operate in the target domain, which drives us to the realm of *domain adaptation* [11, 1, 12, 7, 5, 13, 16].

Learning the optimal feature transformation and an adapted classifier can be seen as two major methods to address the domain adaptation problem. Recent works have been focusing on combining feature transformation and learning adaptive classifiers in a joint max-margin discriminative learning framework [2, 6]. Following this idea, we propose a general domain adaptation method that jointly learns a feature transformation from the source and the target domains to a common subspace and a discriminative classifier in a multi-task framework. This method is inspired by some recent researches *e.g.* bilinear classifiers [9], steerable part models[10].

Our proposal can be applied to general domain adaptation problems and particularly for visual domain adaptation tasks considering that visual data are better modeled in a matrix form rather than in a vector manner [15]. In this work, we focus on cross-domain object detection and apply our method to the state-of-the-art *deformable part-based model* (DPM) which relies on HOG-style features and is trained by latent SVM [4].

The proposed method has connections to some recent works in domain adaptation. Hoffman *et al.* proposed a domain invariant feature representation method, which learns a transformation matrix and the classifier parameter jointly in a max-margin discriminative framework [6]. However, this method is limited by the fact of mapping the data from target domain to the source domain instead of using a common subspace. Duan *et al.* [2] extended the feature replication method of [11] by using a shared feature subspace, thus handling heterogeneous features from different domains. However, the feature replication method increases the computational cost and model complexity. Our method learns a low-rank parameter matrix which reduces the feature dimension, increasing the efficiency in terms of computational cost.

2 Method

We propose to use a bilinear model for visual domain adaptation, which maps the domain specific features to a low dimensional domain invariant subspace and learns a discriminative classifier simultaneously. We start by introducing bilinear classifiers and then we present the mathematical expressions related to the idea of learning cross-domain bilinear classifiers in a multi-task framework. Finally, we show the application of our proposal for domain adaptation of deformable part-based object detectors.

2.1 Bilinear Classifiers

The existing formulations of linear classification typically consider the feature $\mathbf{x}$ in a vector form and the classifier in the form of $f(\mathbf{x}) = \mathbf{w}'\mathbf{x}$, where $\mathbf{w}$ is the model parameter. The bilinear classifiers extend these linear classifiers by using a matrix form of the features. Representing a feature as a matrix $\mathbf{X}$ whose dimensions are denoted by d and c ($\mathbf{X} \in \mathbb{R}^{d\times c}$), the bilinear classifier can be formulated as: $f(\mathbf{X}) = Tr(\mathbf{W}'\mathbf{X})$, $\mathbf{W} \in \mathbb{R}^{d\times c}$, where $Tr()$ is the trace operation. By restricting the rank of $\mathbf{W}$ to be $r \leq \min[d, c]$, we can write the parameter matrix in a decomposed form, $\mathbf{W} = \mathbf{W}'_d\mathbf{W}_c$, where $\mathbf{W}_d \in \mathbb{R}^{r\times d}$ and $\mathbf{W}_c \in \mathbb{R}^{r\times c}$. Thus, the bilinear classifier can be written as: $f(\mathbf{X}) = Tr(\mathbf{W}'\mathbf{X}) = Tr(\mathbf{W}'_c\mathbf{W}_d\mathbf{X})$.

For simplicity, we consider a binary classification problem. Assume we are given a set of training data and label pairs $\{\mathbf{X}_i, y_i\}$, $\mathbf{X}_i \in \mathbb{R}^{d\times c}$ and $y_i \in \{+1, -1\}$. A SVM is often used to optimize the parameters $\mathbf{W}$ of the linear classifier by minimizing the following objective function: $J(\mathbf{W}) = \frac{1}{2}Tr(\mathbf{W}'\mathbf{W}) + C\sum_i \mathcal{L}(y_i, \mathbf{W}, \mathbf{X}_i)$, where $Tr(\mathbf{W}'\mathbf{W})$ is the regularization term, $\mathcal{L}$ is a loss function which penalizes the error on the training samples, and C is used to trade off the loss on the training sample versus the regularization term to control the accuracy. Note that $Tr(\mathbf{W}'\mathbf{W})$ is equivalent to the Frobenius Norm, which can be denoted by $\|\mathbf{W}\|_F^2$. Equivalently, the objective function of a bilinear classifier can be written as: $J(\mathbf{W}_c, \mathbf{W}_d) = \frac{1}{2}\|\mathbf{W}'_c\mathbf{W}_d\|_F^2 + C\sum_i \mathcal{L}(y_i, \mathbf{W}_c, \mathbf{W}_d, \mathbf{X}_i)$. The objective function is not convex but it can be solved by alternative optimization, *e.g.* by coordinate descent as in [9].

2.2 Multi-task Bilinear Classifiers (MT-BL) for Domain Adaptation

Note that the bilinear classifier can be written as $f(\mathbf{X}) = Tr(\mathbf{W}'_c\tilde{\mathbf{X}})$, where $\tilde{\mathbf{X}} = \mathbf{W}_d\mathbf{X}$, and $\tilde{\mathbf{X}} \in \mathbb{R}^{r\times c}$. Thus $\mathbf{W}_d$ is equivalent to a transformation matrix of dimension $r \times d$ and it maps the original

d dimensional feature space to the r dimensional subspace. By using the bilinear classifier, we can map the features from different domains into a common subspace, so that the mapped features are likely to have similar distributions and an *unbiased* classifier can be learned in the common space.

When applying a bilinear classifier for the domain adaptation task, this problem can be formulated in a multi-task learning (MTL) manner where the classifier weight $\mathbf{W}_c$ is the shared parameter among different tasks. For the sake of simplifying the formulations, we consider the adaptation from one source domain $\mathcal{D}^S$ to one target domain $\mathcal{D}^T$. Also we assume the features in each domain have the same dimension, *e.g.* $\mathbf{X} \in \mathbb{R}^{d \times c}$, although our method can also handle heterogeneous features. We assume that the rank of the common subspace is r and define $\mathbf{P}_S \in \mathbb{R}^{r \times d}$ as the transformation matrix for the source domain features and $\mathbf{P}_T \in \mathbb{R}^{r \times d}$ for the target domain ones. We denote by $\mathbf{W}_c \in \mathbb{R}^{r \times c}$ a common subspace classifier parameter matrix and we write the bilinear classifiers for each domain as $f_S(\mathbf{X}) = Tr(\mathbf{W}_c{}'\mathbf{P}_S\mathbf{X})$ and $f_T(\mathbf{X}) = Tr(\mathbf{W}_c{}'\mathbf{P}_T\mathbf{X})$. Hence the goal is to optimize the transformation matrices $\mathbf{P}_S$ and $\mathbf{P}_T$, as well as the common subspace classifier parameter matrix $\mathbf{W}_c$. The multi-task learning optimizes all these parameter matrices jointly with the following objective functions:

$$\begin{aligned} J(\mathbf{W}_c, \mathbf{P}_S, \mathbf{P}_T) &= J_S(\mathbf{W}_c, \mathbf{P}_S) + J_T(\mathbf{W}_c, \mathbf{P}_T), \\ J_S(\mathbf{W}_c, \mathbf{P}_S) &= \frac{1}{2}\|\mathbf{P}_S{}'\mathbf{W}_c\|_F^2 + C_S \textstyle\sum_i^{N_S} \mathcal{L}(y_i, \mathbf{W}_c, \mathbf{P}_S, \mathbf{X}_i^S), \\ J_T(\mathbf{W}_c, \mathbf{P}_T) &= \frac{1}{2}\|\mathbf{P}_T{}'\mathbf{W}_c\|_F^2 + C_T \textstyle\sum_i^{N_T} \mathcal{L}(y_i, \mathbf{W}_c, \mathbf{P}_T, \mathbf{X}_i^T). \end{aligned} \tag{1}$$

Thus the minimization of (1) is a multi-task biconvex problem, where the alternative optimization strategy, *e.g.* coordinate descent in [9] can be applied. By using a re-parameterizing trick, the minimization of (1) can be translated into solving two standard Frobenius-norm-based SVMs. We illustrate the two steps coordinate descent algorithms as follows:

(1) Fixing $\mathbf{P}_S$ and $\mathbf{P}_T$, and optimizing $\mathbf{W}_c$:

By denoting $\mathbf{A} = \mathbf{P}_S\mathbf{P}_S{}' + \mathbf{P}_T\mathbf{P}_T{}'$, $\tilde{\mathbf{W}_c} = \mathbf{A}^{\frac{1}{2}}\mathbf{W}_c$, $\tilde{\mathbf{X}_i^S} = \mathbf{A}^{-\frac{1}{2}}\mathbf{P}_S\mathbf{X}_i^S$ and $\tilde{\mathbf{X}_i^T} = \mathbf{A}^{-\frac{1}{2}}\mathbf{P}_T\mathbf{X}_i^T$, (1) can be written as: $J(\tilde{\mathbf{W}_c}) = \frac{1}{2}\|\tilde{\mathbf{W}_c}\|_F^2 + C_S \sum_i^{N_S} \mathcal{L}(y_i, \tilde{\mathbf{W}_c}, \tilde{\mathbf{X}_i^S}) + C_T \sum_i^{N_T} \mathcal{L}(y_i, \tilde{\mathbf{W}_c}, \tilde{\mathbf{X}_i^T})$.

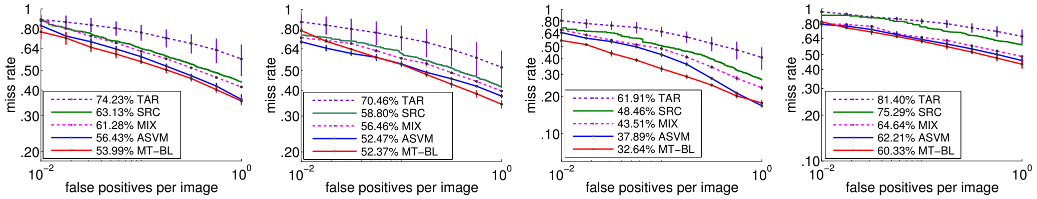
(2) Fixing $\mathbf{W}_c$ and optimizing $\mathbf{P}_S$ and $\mathbf{P}_T$:

Since $\mathbf{P}_S$ and $\mathbf{P}_T$ can be optimized independently, we give the formulation of the optimizing $\mathbf{P}_S$ as an example. Denoting $\mathbf{A} = \mathbf{W}_c\mathbf{W}_c'$, $\tilde{\mathbf{P}_S} = \mathbf{A}^{\frac{1}{2}}\mathbf{P}_S$, and $\tilde{\mathbf{X}_i^S} = \mathbf{A}^{-\frac{1}{2}}\mathbf{W}_c\mathbf{X}_i^{S'}$, the objective function of J_S can be written as $J_S(\tilde{\mathbf{P}_S}) = \frac{1}{2}\|\tilde{\mathbf{P}_S}\|_F^2 + C_S \sum_i^{N_S} \mathcal{L}(y_i, \tilde{\mathbf{P}_S}, \tilde{\mathbf{X}_i^S})$. The same re-parameterizing can be applied to optimize $\mathbf{P}_T$.

Finally, we obtain the optimum $\tilde{\mathbf{W}_c}$, $\tilde{\mathbf{P}_S}$ and $\tilde{\mathbf{P}_T}$. The final classifier for the target domain can be obtained by $f_T(\mathbf{X}) = \mathbf{W}_c\mathbf{P}_T'\mathbf{X}$.

2.3 Domain Adaptation of Deformable Part-based Object Detectors

The bilinear model is well suited for computer vision problems since the visual data are usually in a form of matrix. In this work, we apply the proposed method to the HOG feature based deformable part-based model (DPM) [4]. Given an image $\mathbf{I}$, an object hypothesis in a DPM is defined by $\mathbf{h} = [p_0, p_1, \ldots, p_m]$, where $p_i, (i \in [0, m])$ is the part location. The appearance feature $\Phi_a(\mathbf{I}, \mathbf{h})$ and spatial feature (deformation) $\Phi_s(\mathbf{I}, \mathbf{h})$ are extracted from these part locations. For each part we concatenate all the HOG cells horizontally and as each cell is a f dimensional feature vector, we construct a feature matrix of dimensions $f \times c_i$, where c_i is the number of cells in part p_i. Finally, we concatenate all the part feature matrices into a $d \times c$, $(c = \sum_i^m c_i)$, matrix. For the spatial feature, we keep it in a vector form. Thus, the detection model of DPM can be written as $f(\mathbf{I}, \mathbf{h}) = Tr(\mathbf{W}_a'\Phi_a(\mathbf{I}, \mathbf{h})) + \mathbf{w}_s'\Phi_s(\mathbf{I}, \mathbf{h})$. To cope with the latent variables, *i.e.* $p_i, i \in [1, m]$, the latent SVM is used to optimize the model parameters and refine the part locations alternatively [4]. We incorporate the bilinear model with latent SVM by decomposing the appearance parameter as $\mathbf{W}_a = \mathbf{W}_c'\mathbf{P}$, and the objective function is written as $J(\mathbf{W}_c, \mathbf{P}, \mathbf{w}_s) = \frac{1}{2}\|\mathbf{W}_c'\mathbf{P}\|_F^2 + \frac{1}{2}\mathbf{w}_s'\mathbf{w}_s + C\sum_i^N \mathcal{L}(y_i, \Phi_a(\mathbf{I}_i, \mathbf{h}), \Phi_s(\mathbf{I}_i, \mathbf{h}), \mathbf{W}_c, \mathbf{P}, \mathbf{w}_s)$, where $\mathcal{L}$ is a 0-1 loss function in [4]. We can apply



ETH0 ETH1 ETH2 TUD

Figure 1: Results of adapting a DPM pedestrian detector from synthetic data to work in TUD and ETH pedestrian datasets.

the same multi-task formulas to DPM as in Sec 2.2 for domain adaptation. For example, when **P** is fixed, we optimize $\mathbf{W}_c$ and $\mathbf{w}_s$ in the latent SVM framework, and when $\mathbf{W}_c$ and $\mathbf{w}_s$ are fixed, we optimize **P** in a linear SVM objective function. Since Eq. (1) is none-convex, a good initialization is required for obtaining the optimal. In practice, we calculate the PCA of HOG features from both source and target domain samples, and use the first r eigenvectors as initial value of the **P**. The value of r can be selected by validation as in [17] or subspace disagreement measure as in [5].

3 Experiments and Results

We evaluate our proposal in cross-domain pedestrian detection. Using the MT-BL DPM explained in Sec 2.3, we adapt a pedestrian classifier from a synthetic *virtual-world* dataset [8, 13] to several real-world datasets. In particular, we perfom the adaptation to the three subsets of the ETH [3] dataset (ETH0, ETH1 and ETH2) and the TUD-Brussels dataset (TUD) [14]. We use 200 pedestrians samples (randomly selected) from each target domain and compare with the following baselines:

SRC: Classifier trained with only the source domain samples, *i.e.*, the synthetic ones.

TAR: Classifier trained with only target domain samples, *i.e.*, the real-world ones.

MIX: Classifier trained with source + target domain samples.

ASVM: The regularization based domain adaptive SVM [18], trained with source samples and adapted with the target domain ones.

We perform the evaluation following the Caltech pedestrian detection benchmark and report the average miss rate versus false positive per image. We create five random train/test splits and we average the results across them. The results are shown in Figure 1. Our method produces the highest accuracy on the four target domain datasets demonstrating its ability to achieve a high detection performance and to outperform other competitive domain adaptation approaches.

4 Conclusion

We proposed a multi-task bilinear method to jointly learn a domain-shared feature subspace and a cross-domain discriminative classifier within it. Focusing on visual domain adaptation, we apply this approach to the state-of-the-art deformable part-based detector in order to perform cross-domain pedestrian detection. In particular, we adapt a detector from a virtual synthetic dataset to real-world pedestrian datasets. The results on the Caltech pedestrian detection benchmark show that our method significantly improves the detection performance of a pedestrian detector trained with synthetic data, and outperforms other domain adaptation approaches. In the future, we would like to extend our method to more challenging unsupervised visual domain adaptation tasks to cope with unlabeled data from target domain.

Acknowledgments

This work is supported by the Spanish MICINN projects TRA2011-29454-C03-01 and TIN2011-29494-C03-02, the Chinese Scholarship Council (CSC) grant No.2011611023 and Sebastian Ramos' FPI Grant BES-2012-058280.

References

[1] S. Ben-David, J. Blitzer, K. Crammer, A. Kulesza, F. Pereira, and J.W. Vaughan. A theory of learning from different domains. *Machine Learning*, 79(1):151–175, 2009.

[2] Lixin Duan, Dong Xu, and Ivor W. Tsang. Learning with augmented features for heterogeneous domain adaptation. In *Int. Conf. on Machine Learning*, Edinburgh, Scotland, 2012.

[3] A. Ess, B. Leibe, and L. Van Gool. Depth and appearance for mobile scene analysis. In *Int. Conf. on Computer Vision*, Rio de Janeiro, Brazil, 2007.

[4] P. Felzenszwalb, R.B. Girshick, D. McAllester, and D. Ramanan. Object detection with discriminatively trained part based models. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 32(9):1627–1645, 2010.

[5] B. Gong, Y. Shi, F. Sha, and K. Grauman. Geodesic flow kernel for unsupervised domain adaptation. In *IEEE Conf. on Computer Vision and Pattern Recognition*, Providence, RI, USA, 2012.

[6] J. Hoffman, E. Rodner, J. Donahue, K. Saenko, and T. Darrell. Efficient learning of domain invariant image representations. In *Int. Conf. on Learning Representations*, Arizona, USA, 2013.

[7] Aditya Khosla, Tinghui Zhou, Tomasz Malisiewicz, A. Efros, and A. Torralba. Undoing the damage of dataset bias. In *European Conf. on Computer Vision*, Florence, Italy, 2012.

[8] J. Marín, D. Vázquez, D. Gerónimo, and A.M. López. Learning appearance in virtual scenarios for pedestrian detection. In *IEEE Conf. on Computer Vision and Pattern Recognition*, San Francisco, CA, USA, 2010.

[9] H. Pirsiavash, D. Ramanan, and C. Fowlkes. Bilinear classifiers for visual recognition. In *Advances in Neural Information Processing Systems*, Vancouver, Canada, 2009.

[10] Hamed Pirsiavash and Deva Ramanan. Steerable part models. In *IEEE Conf. on Computer Vision and Pattern Recognition*, Washington, DC, USA, 2012.

[11] H. Daumé III. Frustratingly easy domain adaptation. In *Meeting of the Association for Computational Linguistics*, Prague, Czech Republic, 2007.

[12] K. Saenko, B. Hulis, M. Fritz, and T. Darrel. Adapting visual category models to new domains. In *European Conf. on Computer Vision*, Hersonissos, Heraklion, Crete, Greece, 2010.

[13] D. Vázquez, A.M. López, J. Marín, D. Ponsa, and D. Gerónimo. Virtual and real world adaptation for pedestrian detection. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 2013.

[14] C. Wojek, S. Walk, and B. Schiele. Multi-cue onboard pedestrian detection. In *IEEE Conf. on Computer Vision and Pattern Recognition*, Miami Beach, FL, USA, 2009.

[15] L. Wolf, Hueihan Jhuang, and T. Hazan. Modeling appearances with low-rank svm. In *IEEE Conf. on Computer Vision and Pattern Recognition*, Minneapolis, Minnesota, USA, 2007.

[16] J. Xu, D. Vázquez, S. Ramos, A.M. López, and D. Ponsa. Adapting a pedestrian detector by boosting LDA exemplar classifiers. In *IEEE Conf. on Computer Vision and Pattern Recognition – Workshop on Ground Truth*, Portland, OR, USA, 2013.

[17] J. Yan, X. Zhang, Z. Lei, S. Liao, and S.Z. Li. Robust multi-resolution pedestrian detection in traffic scenes. In *IEEE Conf. on Computer Vision and Pattern Recognition*, Oregon, USA, 2013.

[18] J. Yang, R. Yan, and A.G. Hauptmann. Cross-domain video concept detection using adaptive SVMs. In *ACM Multimedia*, Augsburg, Germany, 2007.