Transferring Boosted Detectors Towards Viewpoint and Scene Adaptiveness

Junbiao Pang, Qingming Huang, *Senior Member, IEEE*, Shuicheng Yan, *Senior Member, IEEE*, Shuqiang Jiang, *Senior Member, IEEE*, and Lei Qin, *Member, IEEE*

***Abstract*—In object detection, disparities in distributions between the training samples and the test ones are often inevitable, resulting in degraded performance for application scenarios. In this paper, we focus on the disparities caused by viewpoint and scene changes and propose an efficient solution to these particular cases by adapting generic detectors, assuming boosting style. A pretrained boosting-style detector encodes *a priori* knowledge in the form of selected features and weak classifier weighting. Towards adaptiveness, the selected features are shifted to the most discriminative locations and scales to compensate for the possible appearance variations. Moreover, the weighting coefficients are further adapted with covariate boost, which maximally utilizes the related training data to enrich the limited new examples. Extensive experiments validate the proposed adaptation mechanism towards viewpoint and scene adaptiveness and show encouraging improvement on detection accuracy over state-of-the-art methods.**

***Index Terms*— Boosting, covariate shift, detector adaptiveness, object detection, transfer learning.**

I. Introduction

OBJECT detection/localization has been extensively studied for more than two decades. Although most prior algorithms have been proposed to detect frontal human faces [30], [37], pedestrians [10], [27], etc., they are believed to be readily extensible to detection of other visual objects, e.g., animals and profile faces. A straightforward extension scheme normally consists of three steps: training examples collection, detection model selection, and detector training. Nevertheless, many object detection tasks are still beyond the capabilities of the state of the art [10], [27], [37]. Even for those nearly solved tasks [37], the high *initial cost*, i.e., the cost to acquire sufficient training examples, may prohibit building practical detection systems.

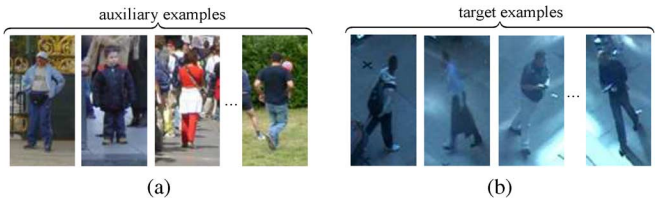


Fig. 1. Data-distribution disparity problem illustrated by examples selected from two different viewpoints. (a) Frontal view. (b) Overhead view.

The high initial cost inherently arises from the fact that most current solutions are based on statistical learning techniques, whereby several thousands of positive examples are typically required to train a detector. For instance, an early face detector uses nearly 1050 face examples [30]. Moreover, detection problems are further complicated by variations of appearance [10], [26]. On the other hand, a considerable number of training images which do not contain any instance of the specific object are also required during the training phase. Negative examples are bootstrapped from these "negative" images to ensure a low false positive rate, e.g., 10^{-4}.

In practice, disparities in data distributions are often inevitable between the training data and test one, possibly resulting from differing viewpoints or scenes. For example, in pedestrian detection, the appearance of a pedestrian may be substantially changed when the capturing viewpoint is altered (see Fig. 1). If we take an example of overhead view as test data while using the detector trained from the frontal view, the detection performance would be seriously hurt. One immediate solution is to retrain the detector with examples recollected from the new viewpoint, incurring again the high initial cost. In addition, specific applications often entail confined scenes, such as a surveillance system with stationary cameras watching a particular region only. This observation suggests the possibility to improve detection performance by adapting a generic detector to the particular scene. Such a specialized detector is expected to perform better than the generic detector in terms of both accuracy and efficiency, since to deal with variable backgrounds tends to increase the complexity of the detector.

We advocate transferring the knowledge residing with visual detectors across viewpoints and scenes. Although examples captured from different viewpoints are generally distinct in appearance, there exists certainly close relationship among them. To determine which part in a generic detector is still useful for

Manuscript received April 14, 2010; revised August 09, 2010 and October 27, 2010; accepted December 21, 2010. Date of publication January 06, 2011; date of current version April 15, 2011. This work was supported in part by the National Natural Science Foundation of China under Grant 61025011, Grant 60833006, and Grant 61035001, the National Basic Research Program of China (973 Program) under Grant 2009CB320906, the Beijing Natural Science Foundation under Grant 4092042, and the NRF/IDM Program under Research Grant NRF2008IDM-IDM004-029. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Erhardt Barth.

J. Pang and Q. Huang are with the Graduate University and the Institute of Computing Technology, Chinese Academy of Sciences, Beijing 100190, China (e-mail: jbpang@jdl.ac.cn; qmhuang@jdl.ac.cn).

S. Yan is with the Department of Electrical and Computer Engineering, National University of Singapore, 11756 Singapore (e-mail: elesyans@nus.edu.sg).

S. Jiang and L. Qin are with the Institute of Computing Technology, Chinese Academy of Sciences, Beijing 100190, China (e-mail: sqjiang@jdl.ac.cn; lqin@jdl.ac.cn).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TIP.2010.2103951

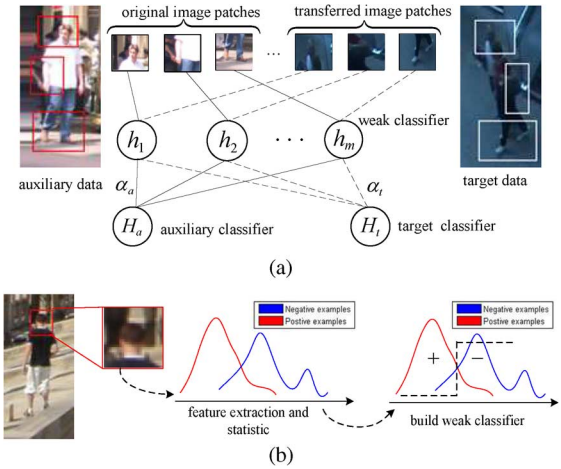


Fig. 2. Schematic of transfer learning across viewpoints or scenes. (a) Transfer learning in boosting: the image patches are transferred from the auxiliary task to the target task. The weight of each image patch is also relearned. Note that the image patches are all normalized into the same size for better visualization. (b) A weak classifier is built upon an image patch, where the parameter of weak classifier and threshold are learned.

a particular case, we utilize a small amount of labeled data captured from the new viewpoints or scenes, called *target-distribution* training data. By comparison, we term the training data for the generic detector as *auxiliary-distribution* data, in view of its potential usefulness for the target task. We transfer the generic detector into the target task by exploiting the relation between the auxiliary data and the target one. This naturally leads to an instance of classic transfer learning [3], [8], [9], [28], [33].

The key underlying argument for our transfer learning is that shared visual features may handle the overall appearance distortion. These shared local features tend to be semantically identical for auxiliary and target examples, but are observed to be at different locations and scales [see Fig. 2(a)]. To establish these features, it is desirable to find correspondences of local features between the different viewpoints or scenes. We start with boosting-style detectors [35], [37] for viewpoint and scene adaptiveness. The reason is that boosted detectors have been successfully applied for detecting various objects, e.g., face [37] and pedestrian [26], [35].

The remainder of this paper is organized as follows. After summarizing the related work, we first review the loss function for the classical boosted detector in Section III and then elaborate on feature shift in Section III-B and on CovBoost in Section III-C. In Section IV, we discuss our methods to transfer the boosted detector across viewpoints and scenes. Extensive experiments on two challenging tasks are presented in Section V. We provide concluding remarks on detection transfer in Section VI.

II. Related Work

Existing possible solutions to the viewpoint and scene adaptiveness problem are partially related to three popular research topics, i.e., multitask learning, semisupervised learning, and transfer learning.

Multitask Learning: Learning for multiple related tasks simultaneously can be advantageous, as compared to learning for these tasks independently [8], [18]. There has been various theoretic work devoted to multitask learning [4], [5]. In computer vision, JointBoost [34] demonstrated that multiple simultaneously trained object detectors behave better than independently learned ones. Recently, Ahmed *et al.* [2] also learned sharing features simultaneously from pseudo (auxiliary) tasks and target tasks with convolutional neural networks (CNNs). Multitask learning can partly solve the deficiency of training examples and possibly improve the performance. However, multitask learning requires that new task has sufficient examples in order to perform simultaneous learning with other related tasks. The high initial cost in the viewpoint or scene adaptiveness is hence inevitable.

Recently, the idea of weighted mixture probabilities [20], [41] suggests to weight distributions among different scenarios, but emphasized storing knowledge in a parametric model [20]. For instance, Taylor expansion of the loss function for auxiliary data is parameterized as the coefficients of weak classifiers [41]. Despite how the possible changes of appearance are handled, [20] and [41] are not designed for either viewpoint or scene adaptiveness. In this work, feature shift is proposed to handle the appearance distortion caused by varying viewpoint or scene.

Semi-Supervised Learning: Another related work is the semi-supervised learning utilizing the unlabeled examples for detectors. One of the popular methods is cotraining [6]—multiple detectors based on independent features are applied to the same unlabeled example, and its label is determined by the highest confidence of detector. To avoid the costly retraining process, the seminal idea [23] has inspired the research in [21] to combine cotraining with an online method [25]. Obviously, cotraining requires different visual cues upon which to build independent detectors. This necessarily brings out the feature design and representation problem, which remains largely open-ended, e.g., bicycles, cars in visual object challenge (VOC)'s tasks [12]. Moreover, the iterative retraining process prohibits practical use. It is also an open problem to apply the generic semi-supervised learning for detectors adaption. For instance, [22] propagated the label information by pair-wise similarity. However, available object examples are often sparse as compared to the large volume of negative examples (the ratio of positive and negative examples can be 1:10 000). How to apply semi-supervised method for adaptation in the case of detection is still vague.

Transfer Learning: Transferring knowledge across related tasks is a known phenomenon in human learning [28]. The related research can be roughly categorized into three classes, according to the level of knowledge transferred. The model-level transfer first estimates the hyper prior of parameters from several related tasks and then transfers this prior to similar tasks, e.g., hierarchical Bayesian models with hyper prior constructed for similar tasks [4], [14], [29], [39]. However, it is generally difficult to model and incorporate priors for discriminative classifiers, which underpin most efficient detection and recognition algorithms, e.g., boosting [16] and support vector machines (SVM) [36] used in detection of faces [37] and human bodies [10], [26], [35]. Second, the data-level transfer instead discovers useful examples from the auxiliary tasks, and uses them along with the target data in a proper strategy. For instance, auxiliary data were used in the covariate shift [32], and the usability of

auxiliary examples was adaptively determined in boosting [9]. Rather than determining the usability of examples by the responses of target classifiers, our approach instead adopts importance sampling mechanism in covariate shift. The third category is the feature-level transfer, which searches for the shared features with satisfactory performance across domains. To uncover these features, one might introduce some related target tasks [3] or learn a distance function which behaves well to transfer knowledge [33]. For instance, Farhadi *et al.* [13] proposed to construct the stable features for recognizing activities from different viewpoints. Comparing with the model-level approach, the data- and feature-level approaches are well suited to transferring knowledge in discriminative models.

The most promising approach for adaptiveness seems to be online boosting [25] (and its variant [17]). With the i.i.d. assumption, the online boosting updates the coefficients of weak classifiers, and requires that weak classifiers have the incremental learning ability. Varying the viewpoint makes the i.i.d. barely hold. Nevertheless, many types of weak classifiers do not have the corresponding incremental versions.

III. Transfer Detector Cross Viewpoints and Scenes

Here, we first review the basic notations for boosting and its applications to object detection. Thereafter, we introduce feature shift and CovBoost for transferring classifiers, respectively [see Fig. 2(a)].

A. Basic Notations and Boosted Detectors

The general approach to object detection is to learn a classifier, which predicts the class label for a subwindow, e.g., 1 for *yes* and -1 for *no*. Within the context of boosting-style detectors, the strong classifier $H(x)$ can be obtained by minimizing the exponential loss $\mathcal{L}$[1]

$$\mathcal{L} = \int_{\Omega} p(x,y)e^{-yH(x)}d(x,y) = \mathrm{E}_{\Omega}\left[e^{-yH(x)}\right] \tag{1}$$

where Ω is the domain of the example-label pair (x,y), which is generated according to the distribution $p(x,y)$, and $y \in \{-1,+1\}$ is the class label of example x. The strong classifier $H(x): x \rightarrow y$ is obtained from a weighted combination of weak classifiers $h_m(x)$

$$H(x) = \sum_{m=1}^{M} \alpha_m h_m(x) \tag{2}$$

where $\alpha_m \in \mathbb{R}^+$ is the coefficient characterizing the importance of the weak classifier $h_m(x)$. The final object detector D is the cascaded strong classifier $H(x)$ [37].

In boosted detectors [37], $h_m(x)$ essentially consists of three elements: 1) location of image patches; 2) parameters of weak classifiers; and 3) decision thresholds for the weak classifiers. Fig. 2(b) illustrates the relation among these elements. For instance, a simple classifier can be obtained by thresholding the Haar feature [37].

[1]Note that, although we focus in this paper on the discrete version of AdaBoost, the proposed approach can be easily extended for other versions of boosting, e.g., RealBoost and LogistBoost.

Let $\mathcal{T}_t = \{(x_i^t, y_i^t)\}_{i=1}^N$ be the target data, where $x_i^t \in \mathcal{X}_t$ is drawn i.i.d. from the target-distribution $p_t(x)$. Let $\mathcal{T}_a = \{(x_j^a, y_j^a)\}_{j=1}^T$ be the auxiliary data, where $x_j^a \in \mathcal{X}_a$ is sampled from the auxiliary-distribution $p_a(x)$.[2] For a particular case of detector transferring across viewpoints, $\mathcal{T}_a$ could represent the examples collected from the horizontal viewpoint, while $\mathcal{T}_t$ could describe the examples collected from other viewpoint (see Fig. 1).

B. Transferring Features by Feature Shift

Denote the location/state of an image patch as $\theta = (l,t,r,b)$, where l,t,r,b are the left-top–right-bottom corner coordinates. Based on the above analysis, the auxiliary state of an image patch should be transferred to the target state θ_t. In other words, the new state θ_t should be determined by using the old state as *a priori* knowledge

$$p(\theta_t|\theta_a) \propto \mathcal{N}(\theta_a, \sigma^2 I) \tag{3}$$

where $\mathcal{N}(\theta_a, \sigma^2 I)$ is a Gaussian distribution with mean θ_a and covariance matrix $\sigma^2 I$. The σ is empirically set to be ten pixels in this work. The Gaussian dependence in (3) means that, in most cases, θ_t deviates slightly from θ_a. As Fig. 2(b) shows, the parameters and the threshold of a weak classifier do not change in this stage. Usually, we can generate L new features based on an old state θ_a. The critical question is how to locate the optimal feature from these L features.

Two strategies are used to locate the optimal feature. One is to directly select from the shifted feature by updating CovBoost in Section III-C1. The other is to average by predicting shifted features according to target data only (see Fig. 3). We will elaborate on these two strategies next.

1) Selecting Shifted Features: Feature shift first generates a new enlarged feature pool, and then the optimal feature is selected by stage-wise optimization in CovBoost. Advantages of this strategy are twofold: simultaneous updating CovBoost and supplying more features for boosting than the following averaging strategy.

2) Averaging Shifted Features: Averaging shifted features uses the auxiliary state as an initial guess to predict the target state, according to the target data only. For a clear presentation, we use $h^a(x)$ instead of $h_m^a(x)$ to represent the mth weak classifier. The state θ_t can be computed by estimating the probability $p(\theta_t|\mathcal{T}_t, \mathcal{T}_a)$ in terms of Bayesian inference. However, the conditional probability $p(\theta_t|\mathcal{T}_t, \mathcal{T}_a)$ cannot be computed directly. Hence, conditional independence between $\mathcal{T}_a$ and $\mathcal{T}_t$ is adopted to simplify $p(\mathcal{T}_t, \mathcal{T}_a|\theta_t)$ as

$$p(\mathcal{T}_a, \mathcal{T}_t|\theta_t) = p(\mathcal{T}_a|\theta_t)p(\mathcal{T}_t|\theta_t). \tag{4}$$

Based on the Bayes rule and (4), we have $p(\theta_t|\mathcal{T}_t, \mathcal{T}_a) \propto p(\mathcal{T}_t|\theta_t)p(\theta_t|\mathcal{T}_a)$. Further, we have

$$\begin{aligned} p(\theta_t|\mathcal{T}_t, \mathcal{T}_a) &\propto p(\mathcal{T}_t|\theta_t)p(\theta_t|\mathcal{T}_a) \\ &= (\mathcal{T}_t|\theta_t) \int p(\theta_t|\theta_a)p(\theta_a|\mathcal{T}_a)d\theta_a \end{aligned} \tag{5}$$

[2]Hereafter, the notation t and a generally represent the target and the auxiliary data, respectively.

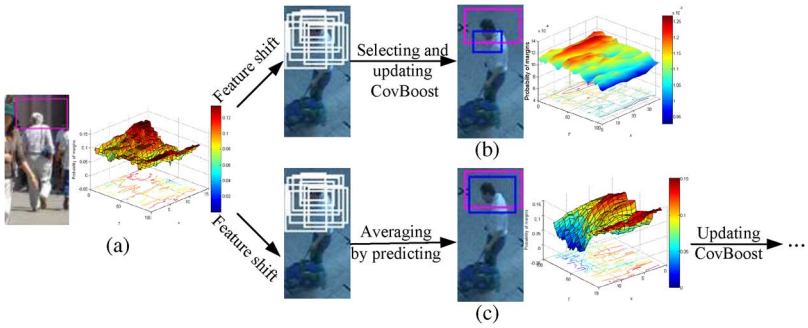


Fig. 3. Feature shifts from the old (pink, online version) state to a new (blue, online version) state by different strategies. The margin gain (6) of the weak classifier on the target data is plotted in (a)–(c). Further comparing the maximum points of the margin gain map in (a)–(c), we can see that the feature often shifts to the local maximum in the margin gain map. Note that the map describes the discrimination capability of weak classifiers. (a) Old state. (b) Target by selecting strategy. (c) Target state by averaging strategy.

where

$$p(\mathcal{T}_t|\theta_t) \propto \frac{1}{\sum_i \exp\left(-y_i^t h^a\left(x_i^t\right)\right)} \tag{6}$$

is the margin gain of the weak classifier $h^a(x)$ on the target data, and $p(\theta_a|\mathcal{T}_a)$ is the probability that the weak classifier occurs at the old state θ_a. $\exp(-y_i^t h^a(x_i^t))$ in (6) is a variation of margin [15], where $y_i^t h^a(x_i^t)$ is positive if and only if $h^a(x_i^t)$ correctly classifies a example. Therefore, the essence of (5) is to search for the target status θ_t by following the largest margin criterion.

The optimal target status θ_t can be estimated via uniform sampling [24] as follows. A set of samples $\{s_l^a\}_{l=1,\ldots,L}$ are generated by repeating the old state θ_a. After the state s_l^a shifts to s_l^t with (3), the state s_l^t is associated with the weights $\pi_l^t \propto p(\mathcal{T}_t|s_l^t)p(s_l^t|s_l^a)$ with $\sum_l^L \pi_l^t = 1$. Monte Carlo is used to approximate the optimal target state θ_t as the expectation $\hat{\theta}_t = \sum_{l=1}^L s_l^t \pi_l^t$. In this paper, we set L to be 50.

3) Discussions on Feature Shift: Pose variation is one of the main difficulties for visual object detection. To handle this problem, multiple instance learning (MIL) [38] uses a set of instances to encode the variation of appearance, and then trains a detector to discover the aligned instances. In this regard, feature shift uses a set of image patches to encode the variation at the feature level. Furthermore, feature shift is a general framework to achieve the feature-level transfer; more types of transforms (e.g., affine or perspective) can be easily incorporated into our system.

C. Covariate Boost

Although the auxiliary distribution $p_a(x)$ is generally different from the target distribution $p_t(x)$ ($p_a(x) \neq p_t(x)$), the conditional probability distribution can be considered to be equal, namely $p_a(y|x) = p_t(y|x)$. *Covariate shift* [32] therefore can be applied to utilize the auxiliary distribution $p_a(x, y)$. Applying covariate shift into the exponential loss (1), we have the covariate loss

$$\widetilde{\mathcal{L}} = \mathrm{E}_{\mathcal{T}_t}\left[e^{-yH_t(x)}\right] + \mathrm{E}_{\mathcal{T}_a}\left[\lambda e^{-yH_t(x)}\right] \tag{7}$$

where $\lambda = p_t(x, y)/p_a(x, y)$ is the ratio of the target and auxiliary density. As illustrated in Fig. 4, there are three pivot points, A, B, and C, where $\lambda \cong 1$. At the white region (B-C), $\lambda \geq 1$,

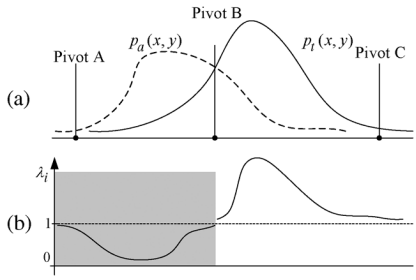


Fig. 4. Analysis of the meaning for ratio λ in (7). (a) The dotted line represents the distribution of auxiliary data $p_a(x, y)$, whereas the solid line represents the distribution of target data $p_t(x, y)$. (b) The value of λ for every example is shown with respect to the distribution disparity.

whereas at the gray region (A-B) $\lambda \leq 1$. The more λ approaches 1, the more similar are the distributions; otherwise, the distributions are more distinct. It should be noted that λ does not mean the usability of auxiliary examples, but indicates the disparity between the target and the auxiliary distribution.

Essentially, the second term of (7) uses the $p_a(x, y)$ as the proposal probability in importance sampling to reuse the auxiliary data $\mathcal{T}_a$. We reformulate the density ratio λ with the conditional probabilities by using the Bayes rule

$$\begin{aligned}\lambda &= \frac{p_t(x,y)}{p_a(x,y)} \\ &= \frac{p(x,y|t)}{p(x,y|a)} \\ &= \frac{p(t|x,y)p(x,y)}{p(t)} \frac{1}{\frac{p(a|x,y)p(x,y)}{p(a)}} \\ &= \frac{p(t|x,y)p(a)}{p(a|x,y)p(t)}.\end{aligned} \tag{8}$$

It is natural to presume that human has equal possibility to be observed in different view angles or scenes; thus, the probability of target view angle $p(t)$ and of auxiliary viewpoint $p(a)$ are assumed to be equal. For other adaptiveness tasks, $p(a)$ and $p(t)$ should also be determined by the domain knowledge. Therefore, (8) can be estimated by the ratio of conditional probabilities, $p(t|x, y)$ and $p(a|x, y)$ (measuring the likelihood an example (x, y) belonging to the target data $\mathcal{T}_t$ or the auxiliary data $\mathcal{T}_a$,

TABLE I
COMPARISON AMONG DIFFERENT METHODS. NOTE THAT $\mathrm{Loss}(x_j^a, y_j^a)$ IS THE LOSS OF A SINGLE EXAMPLE, WHICH CORRESPONDS TO $e^{-yH(x)}$ IN (1)

Method	Loss function	Comments
Multi-task learning	$\mathcal{L}_a+\mathcal{L}_t$	Learn all tasks simultaneously and unbiasedly.
Method [20], [41]	$(1-\lambda)\mathcal{L}_a+\lambda\mathcal{L}_t$ [20] ($\mathcal{L}_a+\lambda\mathcal{L}_t$ [41])	The $\lambda(0\leq\lambda\leq1)$ controls the degree of adaptiveness. If $\lambda = 0$, there is no adaption process. If $\lambda = 1$, only target data is used to learn. The optimal λ can be estimated via cross-validation technique. $\mathcal{L}_a+\lambda\mathcal{L}_t$ [41] can be easily changed into $(1-\lambda)\mathcal{L}_a+\lambda\mathcal{L}_t$ [20], thus λ in [41] has similar meaning.
Our method	$\sum_i \lambda_i \mathrm{Loss}(x_i^a, y_i^a)$ $+\mathcal{L}_t$	The $\lambda_i(0<\lambda_i<+\infty)$ acts as example "selector". If $\lambda_i \approx 0$, (x_j^a, y_j^a) will be useless; otherwise, (x_j^a, y_j^a) will contribute to adaption. The more larger λ_i means that (x_j^a, y_j^a) be more useful for classifier adaption. λ_i can be estimated via (8).

respectively). Here, we model these conditional probabilities as the logistic functions

$$p(t|x,y)=\frac{1}{1+e^{-yH_t(x)}} \quad p(a|x,y)=\frac{1}{1+e^{-yH_a(x)}} \tag{9}$$

where $H_a(x)$ is trained on the auxiliary data $\mathcal{T}_a$, and $H_t(x)$ is adaptively trained with examples from both the auxiliary data $\mathcal{T}_a$ and the target data $\mathcal{T}_t$. The λ has an analytical form

$$\lambda=\frac{1+e^{-yH_a(x)}}{1+e^{-yH_t(x)}}. \tag{10}$$

If $y=1$ in (10), it means that the positive auxiliary examples are used for viewpoint adaptiveness. On the contrary, if $y=-1$ in (10), the negative auxiliary examples from new scenes are utilized for scene adaptiveness. That is, our method can treat both the viewpoint and the scene adaptiveness in a unified framework. The covariate loss (7) can be further written as

$$\widetilde{\mathcal{L}}=\mathrm{E}_{\mathcal{T}_t}\left[e^{-yH_t(x)}\right]+\mathrm{E}_{\mathcal{T}_a}\left[\frac{1+e^{-yH_a(x)}}{1+e^{yH_t(x)}}\right]. \tag{11}$$

The loss $\widetilde{\mathcal{L}}$ consists of two different data sources: the auxiliary-data $\mathcal{T}_a$ and the target-data $\mathcal{T}_t$. Rather than these weighted mixture training [20], [41], CovBoost weights *every* auxiliary example (x_j^a, y_j^a). Table I further summaries the difference among the multitask learning, approaches [20], [41] and our method. By comparing the loss functions in Table I, our approach is distinct in methodology to handle the classifier adaptiveness problem.

1) Optimizing the Covariate Loss: Our method adopts the stage-wise optimization method. We select or update the weak classifier $h_m^t(x)$ by minimizing the first-order approximation of covariate loss (11) as

$$\mathrm{E}_{\mathcal{T}_t}\left[e^{-yH_t(x)}yh_m^t(x)\right]+\mathrm{E}_{\mathcal{T}_a}\left[\frac{\left(1+e^{-yH_a(x)}\right)e^{yH_t(x)}}{\left(1+e^{yH_t(x)}\right)^2}yh_m^t(x)\right]. \tag{12}$$

After the weak classifier $h_m^t(x)$ is updated, the optimal coefficient α_m^t for $h_m^t(x)$ can be solved by minimizing covariate loss (11) with two kinds of methods, i.e., linear searching or gradient based method.

The efficiency can be further improved by recording the weight $D_t(x^t) = e^{-y^t H_t(x^t)}$ for the target data and $D_a(x^a) = e^{y^a H_t(x^a)}$ for the auxiliary data iteratively. Moreover, the mth weak classifier $h_m^t(x)$ and α_m^t can be computed by using only the weight D_t and D_a.

2) Discussions on CovBoost: It may be argued that the target and the auxiliary data can be considered as the mixture distribution, and then classifiers can be trained based on mixture estimation. On the other hand, classifiers could also be directly trained on the target data. We refer to these thoughts as mixture training and single training respectively. For mixture training, more variation of appearance from new viewpoints will enlarge the intra-class variability, which may damage the ability of traditional classifiers. Second, training on a small number of target examples locally often causes a serious overfitting problem. A comparison among mixture training, single training, and CovBoost is presented in Section V-C2.

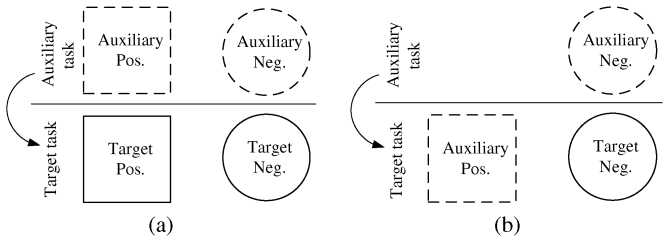


Fig. 5. Relation of the different types of training data between two tasks. Note that the dotted rectangles or circles describe the examples from auxiliary tasks, whereas the solid rectangles or circles represent the examples from target tasks. (a) Angle-scene adaptiveness. (b) Scene adaptiveness.

IV. TRANSFERRING BOOSTED DETECTOR ACROSS VIEWPOINT AND SCENE

In this study, we apply the above transfer algorithm into the cascaded detector proposed by Viola and Jones [37] for two visual tasks. As discussed in Section III-C, the transfer learning algorithm can easily be applied to two scenarios: viewpoint and scene adaptiveness. In essence, the difference between these two tasks lies at utilizing different types of training examples (see Fig. 5). In the following, we will elaborate on these adaptiveness tasks, and consider how to update the other parameters of weak classifiers.

A. Viewpoint-Scene Adaptiveness

Viewpoint-scene adaptiveness means that transferring generic detectors to a specific scenario with new viewpoint

and scene. This task covers a series of instances, for example, changing the detectors trained for the frontal viewpoint into the overhead viewpoints in surveillance. After the viewpoints are changed, a detector is often fixed with viewpoints, and monitors a special scene. One of the main challenges of this task is the transferred detectors should efficiently recognize these instances in new viewpoints when the appearance of objects is seriously changed. A small number of target data barely covers all possible appearances of objects, and hardly guarantees the generalization ability. Therefore, it is necessary to "borrow" examples from the auxiliary data in Fig. 5(a). The viewpoint-scene adaptiveness process is described in Algorithm 2.

Algorithm 1. Transfer Boosted Detector Across Angle-Scene

1: **Given**: The target positive data $\mathcal{T}_t^+$ and the initial target negative data $\mathcal{T}_t^-$ extracted from new viewpoint and scene, respectively. The learned auxiliary cascaded detector $D_a = \{H_a^1(x), \ldots, H_a^K(x)\}$, where $H_a^k(x)$ is kth stage auxiliary classifier and the auxiliary positive examples $\{\mathcal{T}_a\}$ for every stage.
2: **For** $k = 1, \ldots$
— Using feature shift to generate a new feature pool for CovBoost.
— Updating the auxiliary $H_a^k(x)$ into target $H_t^k(x)$ via CovBoost, where the parameters or the thresh of weak classifiers are also updated.
— Bootstrapping the hard negative examples from the auxiliary negative images set for next stage classifier.
End For
3: **Output**: The transferred angle-scene detector $D_t = \{H_t^1(x), \ldots, H_t^k(x), \ldots\}$.

B. Scene Adaptiveness

This task means that transferring generic detectors to particular scenes without viewpoint changes. There are wide applications for scene adaptiveness, for instance, applying detectors from the outdoors into the indoors scenes, adapting detectors from generic scenes into special application scenes, e.g., teleconference room [41]. Intuitively, this task only requires detectors to aggressively reject more negative examples from the *limited* scenes than auxiliary detectors. Therefore, a small set of images without instances of the particular objects are supplied as target negative examples. Similar to viewpoint-scene adaptiveness, the small number of target negative examples can not cover all scenes in novel scenarios. The auxiliary negative examples are also used to increase the generalization ability of detectors in Fig. 5(b). In this paper, we simply achieve scene adaptiveness by appending new classifiers trained by CovBoost onto the auxiliary detector (see Algorithm 3). The false positives generated by step-3 in Algorithm 3 are too "hard" for the initial target detector, and thus "selecting strategy" is used to generate a feature pool for transferring the learned auxiliary classifier $H_a^K(x)$.

Algorithm 2. Transfer Boosted Detector Across Scenes

1: **Given**: The last stage auxiliary negative data $\mathcal{T}_a^-$, the auxiliary positive data $\mathcal{T}_a^+$, the auxiliary cascaded detector $D_a(x) = \{H_a^1(x), \ldots, H_a^K(x)\}$, where $H_a^k(x)$ is k-th stage auxiliary classifier. The target negative images sampled from new scenes.
2: Initializing the target detector D_t, $D_t \leftarrow D_a$.
3: Bootstrapping the hard negative examples $\mathcal{T}_t^-$ for the target negative images by the detector $D_a(x)$.
4: Using "selecting strategy" in feature shift to generate a feature pool from the last staged classifier $H_a^K(x)$.
5: Training a target classifier $H_t(x)$ using the hard negative examples $\mathcal{T}_t^-$, auxiliary negative examples $\mathcal{T}_a^-$ and auxiliary positive examples $\mathcal{T}_a^+$ via CovBoost.
6: Appending the target classifier $H_t(x)$ onto $D_t(x)$, $D_t(x) \leftarrow \{D_t(x), H_t(x)\}$.
7: Return to Step3, until the target negative images are corrected classified.
8: **Output**: the transferred scene detector $D_t = \{H_a^1(x), \ldots, H_a^K(x), \ldots, H_t(x)\}$.

C. Updating Weak Classifier

Feature shift only updates the location of weak classifiers, whereas the threshold and parameters of weak classifier could also be updated. There are two straightforward strategies to deal with these elements: 1) online updating and 2) retraining weak classifiers. To our best knowledge, not all types of weak classifiers have corresponding online versions. Although retraining weak classifiers is somewhat time-costing, we only adopt it to update the threshold without adjusting the parameters of weak classifiers. Taking the single-threshold weak classifier [37] as an example, the threshold of a weak classifier can be simply recomputed to maximize (12). It is also easy to extend the single-threshold weak classifier into the decision-tree-based weak classifier [16].

V. Experiments

The proposed algorithm has been thoroughly tested on both the synthetic and the real datasets. In Section V-A, we show that the ratio of conditional probabilities is efficient to describe the disparity of the different distributions. In Section V-B, we evaluate its ability on pedestrian detection for two tasks, and also verify its ability on another visual object, i.e., human face.

A. Experiments on Synthetic Data

To give an intuitive illustration of CovBoost, Fig. 6 shows 2-D synthetic data with the distribution disparity to emphasize the advantage of using auxiliary data. The synthetic data are first presented by [32], and consists of two parts, i.e., auxiliary data and target ones. Fig. 6 shows that we are considering a non-i.i.d. problem: the auxiliary data is totally different from the target ones.

We generate 3000 auxiliary data and 20 target data from the distributions in Fig. 6(a). Two naïve strategies, mixture training and single training are evaluated. As expected, the decision plane output by CovBoost is more accurate. Also notice that single training and mixture training overfit to the target and the auxiliary data respectively. One can immediately see that CovBoost can properly use auxiliary data to improve performance.

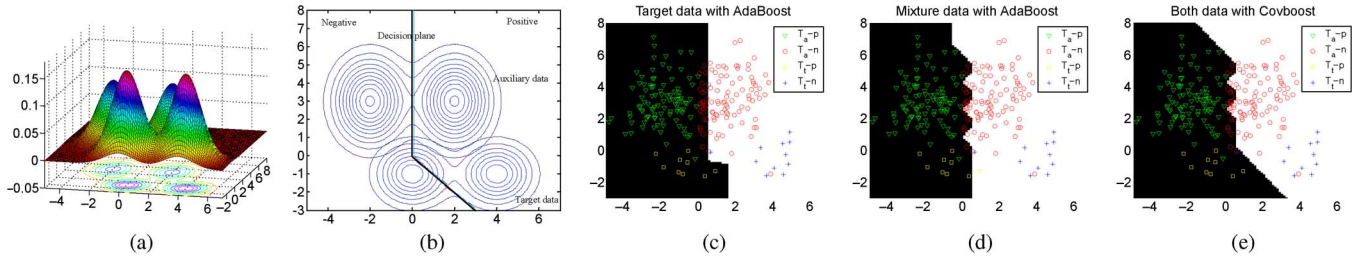


Fig. 6. Different strategies are applied to utilize the auxiliary data. In boosting, only 100 decision stumps are supplied. (a), (b) Synthesis data and the corresponding decision plane. (c)–(e) Legend "T_a − p", "T_a − n", "T_t − n", and "T_t-p" represent the positive auxiliary data, the negative auxiliary data, the negative target data and the positive target data respectively. Note that only 30% auxiliary data are plotted for better illustrating the decision plane. For more details, please refer to Section V-A. (a) Synthetic data. (b) Decision plane. (c) Single training. (d) Mixture training. (e) CovBoost.

To test the ability of estimating the λ in (8), a comparison is done between the state of the art, i.e., importance weighted cross validation (IWCV) [32], and our method. IWCV needs a precondition that the probability of auxiliary data $p_a(x, y)$ and target data $p_t(x, y)$ are already known or estimated, and further applies the weighted cross validation to select best λ. However, collecting a proper scale of validation data for cross validation somewhat falls into the high initial cost problem in the terms of adaptiveness. In contrast, our method does not require the probability $p_a(x, y)$ and $p_t(x, y)$ to be estimated, but the conditional probability of auxiliary data $p(a|x, y)$. Following the test procedure described in [32] on this dataset, we sample 20 target data (10 for positive and 10 for negative examples) for transferring. 0.156 ± 0.041 error rate is achieved with our method, while IWCV reports 0.108 ± 0.026 by adaptive importance weighted Fisher linear discriminant analysis as classifier. From this toy data, we can conclude that the ratio of conditional probabilities is an efficient approach to estimate the disparity of data, and achieves comparable performance with IWCV.

B. Experiments on Real Data

Two diverse datasets are used to evaluate the effectiveness of the proposed algorithm in transferring detectors across viewpoints and scenes. The first dataset is obtained from the PETS 2007 [1], which is captured at a resolution of 720×576 pixels from the real environment without constraints. For the target training set, 220 target positive examples (with reflection) is randomly extracted from the *Dataset S8 view3* and normalized into the size of 128×64 pixels. The *Dataset S7 view3* is labeled every 10 frames as the test set, which amounts to a total of 300 frames with 973 pedestrian instances. These video sequences contain many challenges which are representative in real-life cases: the pose near the camera would be changed intensively with comparison to the ones from the front-view; different actions and moving directions produce diverse poses; unconstrained video streams exhibit a much lower quality than their photographed counterparts [10]. Above factors would enlarge the degree of intra-category variability, and thus increase difficulty for detectors. The auxiliary data are borrowed from the INRIA pedestrian dataset [10]. Fig. 1 shows some samples used in the training stage. Although some examples have similar appearance as the data from the frontal view, mostly the human bodies in the target data is transformed into a slant direction by the new viewpoint.

The second dataset, ETHZ dataset [11], is recorded at a resolution of 640×480 pixels, using a stereo pair of cameras mounted on a children stroller. Only the videos captured with the left camera are used for training and test in our scene adaptiveness task. We sample a small number of target negative images (87 image patches) from the training sequences. The INRIA pedestrian dataset [10] is also used to train the auxiliary detector. The training sequence shows a walk over a fairly busy square on a cloudy day. The first test sequence is taken under a similar weather condition, strolling on a sidewalk, whereas the second sequence shows a stroll over a busy square in the shadow. Here, we ignore the challenges of appearance itself, for instance, partial occlusions between pedestrians, large range of scales of human, multitude of viewpoints, etc. The scenes in two test sequences even pose several main difficulties: a large number of trees and dust bins; reflections from shopping windows; bad weather condition resulting in low contrast; video streams suffering from slight motion blur or some times missing contrast. As a comparison, above difficulties make itself different from the INRIA dataset. Although ETHZ dataset is not originally built for scene adaptiveness, this work adopts it to this task for its disparity from INRIA data.

The false positives per image (FPPI) is used as the evaluation metric. A detection is to be counted as correct, only if it has to overlap with an annotation by more than 50%, using the intersection-over-union measure [12].

C. Experiments on PETS2007 for Viewpoint-Scene Adaptiveness

1) Systematic Experiments: The experiments in this subsection are performed to determine the optimal choices in feature shift. Therefore, an analysis is done to study the efficacy of feature shift strategies: averaging strategy, selecting strategy and without feature shift (WFS). In this experiment, the thresholds of weak classifiers are not changed.

As expected, WFS has the worse result on accuracy in Fig. 7. This is mainly due to that the intensive change of appearance is too difficult to WFS. Variation resulting in large intra-class variability becomes a major problem in the angle-scene adaptiveness task. Further, the averaging strategy performs better than WFS, but worse than the selecting strategy as shown in Fig. 7. Although the averaging strategy is explicitly designed to account for viewpoint change, the reasons for its inferior accuracy may be twofold: 1) only target data is utilized to predict the states of weak classifiers and 2) a small number of feature pool is built by the averaging approach. The latter also confirms the

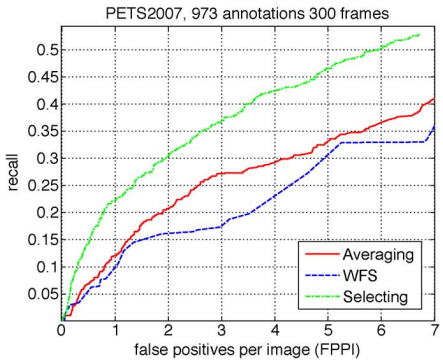


Fig. 7. Effect of feature shift methods with HOG feature.

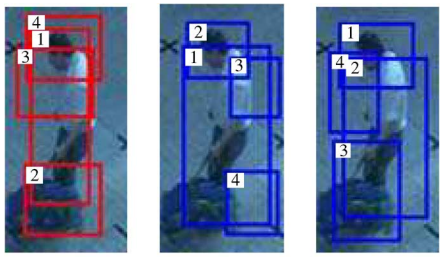


Fig. 8. Illustration of the four features selected by the feature shift. (a) Features selected by the auxiliary classifier. (b) Features selected by the averaging strategy. (c) Features selected by the selecting strategy. The tags of rectangles show the sequence of features selected in boosting iteration.

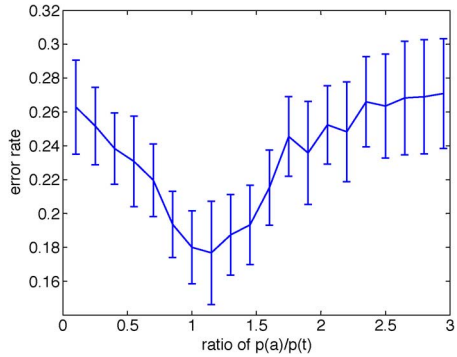


Fig. 9. Impact of different $p(a)/p(t)$ on the classification ability. The means are the averages of 10 random repeats, as well as their standard deviations. The accuracy is evaluated as: #the miss-classified/#total examples.

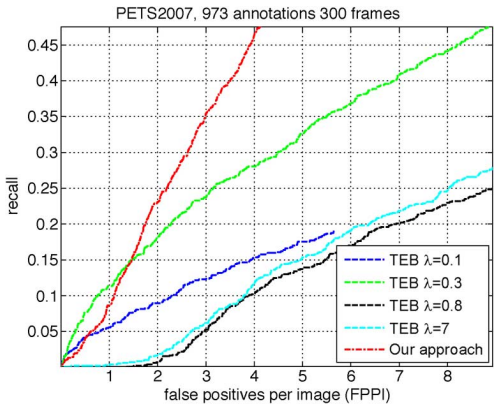


Fig. 10. Comparison with Taylor-based method [41].

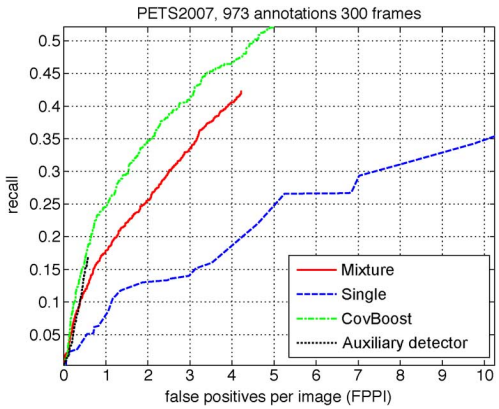


Fig. 11. Comparison with two naive thoughts.

fact: the number of weak classifiers should be large enough to obtain better performance [26], [35], [37] in boosted detectors.

To understand the correspondence among shifted features, we build a auxiliary classifier with 4 features, and shift them according to the different strategies. It can be seen that the position and the selected sequence are different. For example, the first feature in Fig. 7(b) corresponds to the first feature in Fig. 7(a), while the fourth feature in Fig. 7(a) turns into the first feature selected by selecting strategy. These difference empirically shows that the auxiliary data can influence the boosting iteration in CovBoost.

The ratio of $p(a)/p(t)$ is also empirically evaluated to verify our assumption in (9). In this experiment, 700 positive and 1000 negative examples are extracted as test data from PETS2007; only 100 positive examples and 500 negative examples are used as target data for training. It can been seen that the region [1.0,1.3] approximates at the bottom of the curve in Fig. 9. Although the minimal error is not exactly achieved at 1.0, it is still valid to assume that human has the same probability to be observed under different view angles.

2) Comparative Experiments: With the possible choice motivated in the previous subsection, we now apply the proposed system to the challenging test sequence—*Dataset S7 view3* in PETS2007 is a surveillance video watching the intersection between two alleyways from the overhead viewpoint. By using this test set, we compare our system to the state of the art [41] and several naïve thoughts.

First, we implement a version of our system based on Haar feature [37], and carry out a comparison between the Taylor expansion based method (TEB) [41] and our system. Because TEB is the most related work for detector adaptiveness, and uses Haar feature. For a fair comparison, we implement rigorously the "Direct labels" method in [41], and tune all the listed values of the parameter λ (the middle value is adopted in our experiment). It should be noted that TEB only updates the coefficient α_m^t of weak classifier by optimizing the hybrid loss in Table I.

The performance of the auxiliary detector is not plotted in Fig. 10, because the auxiliary detector rejects all image patches as negative examples in this experiment. In comparison to TEB, near 5% gain is achieved by our approach at $\mathrm{FPPI} = 2$. The possible reasons for the success of CovBoost will be further discussed in Section V-C4.

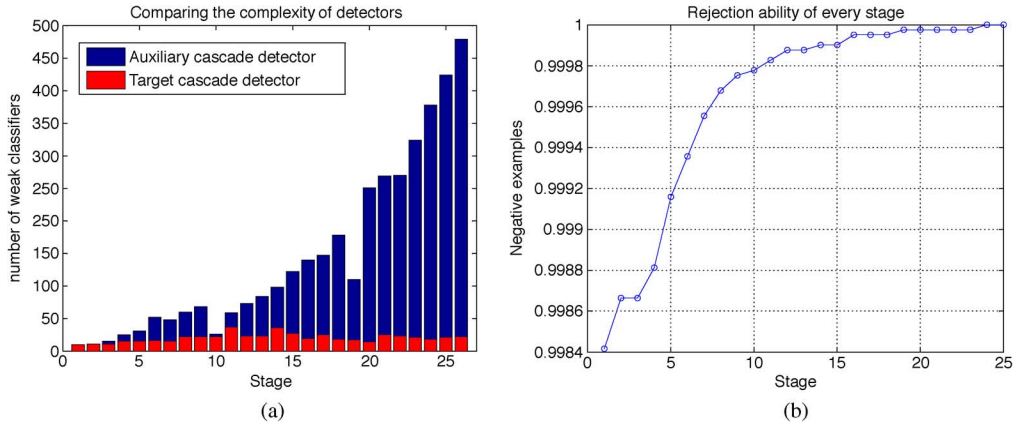


Fig. 12. Computational efficiency of transferred detector. (a) Comparing the complexities of auxiliary and target detectors. The ratio of total number of weak classifiers between the auxiliary detector and the target one approximates 7:1. (b) Accumulated rejection ability of detectors.

In next experiment, we implement another version of our system based on histogram of gradients (HOG) feature [10]. Because the HOG is the widely used descriptor for human detection [10], [12], [26]. Although a thorough comparison is done among the single training, mixture training and our approach on synthesis data, Fig. 11 gives another comparison in real data. From Fig. 11 we can further analyze how the auxiliary data affect the performance and how proper using the auxiliary data can improve the generalization power.

As illustrated in Fig. 11, the single training approach faces a serious overfitting problem, which produces many false positives and false negatives than those using auxiliary data, i.e., mixture training and CovBoost. This comparison results show that the auxiliary data indeed contains reusable examples for the related tasks, and incorporating the auxiliary data can retain the generalization ability of detectors.

A comparison between mixture training and CovBoost shows the importance of proper usage of auxiliary data. Intuitively, the small number of target data tend to be overlooked by AdaBoost in mixture training, due to the distribution disparity, as well as the large ratio between the number of auxiliary and target data. However, weighting every examples can acts as an example selector to emphasize the examples in the vicinity of target data, but suppress the ones far from target data. Thus, the effect of the imbalance between the target and the auxiliary data is adaptively countered.

3) Complexity of Transferred Detector: To analyze the complexity of transferred detectors, we measure the "complexity" of boosted detectors—the sum of weak classifiers in a boosted classifier. The cascaded detectors distribute the computation process of a detector among the rejection stages. Therefore, a smaller number of weak classifiers in a stage implies faster detection. Fig. 12(a) shows that the number of weak classifiers of the transferred detector is far smaller than that of the auxiliary detector. This means that, if an example goes through the whole cascaded detector, the transferred detector can still enjoy approximately seven times faster detection speed than the original auxiliary detector.

Moreover, the detection speed is also closely related with the "rejection ability" in a cascaded detector, which measures the number of negative examples rejected by the classifier in every stage. The earlier the cascade rejects more negative examples, the faster the detection is. This is obvious since the classifiers in latter stages often consist of more weak classifiers. For the limited scenes, the fist stage can efficiently reject nearly 99% negative examples, and as a result, the left negative examples are difficult to the rest of stages. Consequently, the transferred detector costs ten stages to reject 0.1% hard negative examples in Fig. 12(b).

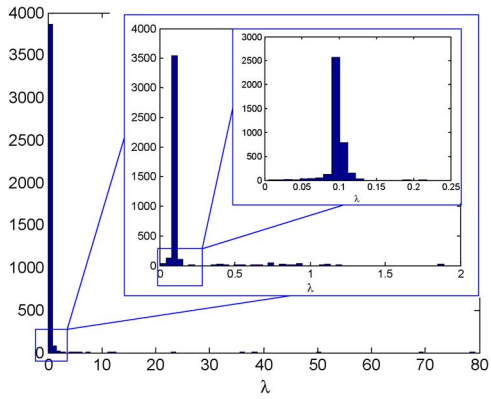


Fig. 13. Distribution of λ in CovBoost, where a Haar-like feature is used. For better visualizing the distribution of λ, a series of multiple scaled histograms is plotted. The mean of λ_i is 0.334 with std 3.115.

Correlating the accuracy and the complexity of the transferred detector, it is necessary and possible to transfer a detector to a scenario with new viewpoints. Because transferred detectors often obtain faster detection and more accurate performance than auxiliary detectors.

4) Distribution Disparity: Fig. 6 has already given a glimpse of the distribution disparity problem, which is successfully handled by CovBoost. Further, an investigation is done to study why weighting every examples is better than weighting the whole distribution approach in terms of improving the generalization ability of detectors.

Fig. 13 shows the distribution of λ of every example at different scales. It is obvious that λ for most auxiliary examples concentrates around 0.1, which means that, auxiliary data does contribute to detector adaptiveness. Further checking the target data and auxiliary one in Fig. 1, there are still some common points in shape: these data all have a rectangle-like shape. This

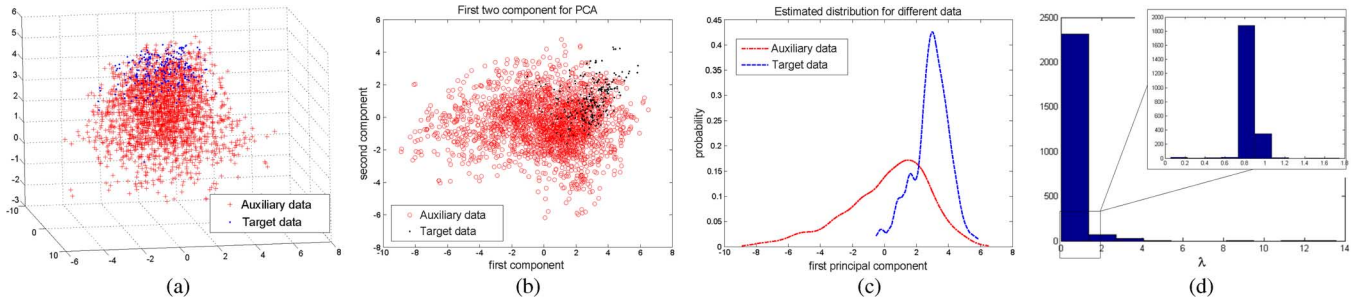


Fig. 14. Low dimension distribution of the auxiliary data and target data. The probabilities in (c) are obtained by applying Parzen window on the first component. (a) 3-D. (b) 2-D. (c) Estimated probability. (d) Distribution of λ.

may explain why most of examples in auxiliary data still contributes to the target detector, albeit in a limited way. The large standard deviation (std), 3.1, indicates that a small number of auxiliary examples live in pivot B-C in Fig. 4. In other words, in terms of handling distribution disparity, CovBoost gives higher weight to these examples with visual similarity to target data. Therefore, determining the λ in an adaptive approach gives more flexibility and accuracy than weighted probability approach [20], [41].

Fig. 14 further shows the low-dimension distribution of two different data with principal component analysis (PCA), from which we can understand the relation between these data, and the possibility to reuse the auxiliary data. The HOG feature first is built by cells with 4×4 pixels, and then every example is densely sampled with a block 2×2 cell. There are 188 blocks with 6768 $(188 \times 36 = 6768)$ dimensions for every example, which is similar to the approach in [10] to represent human. Dense sampling at different scales can then efficiently reduce the variation problem among examples. Latter, PCA is applied on these 2636 examples (2416 auxiliary positive data+220 target positive data) into 2-D and 3-D space, respectively.

The distribution disparity is very serious: target data only has a partial overlap with the auxiliary data [see scattered points in Fig. 14(a) and (b)], e.g., in Fig. 14(b), some examples are far away from the auxiliary data, making AdaBoost easily treat themselves as noise, and thus causing AdaBoost to be sensitive in mixture training [7]. Nevertheless, the self-tuning λ turns these sparse target data as landmarks to accord more attention to vicinal auxiliary examples and to endow low weight to the auxiliary examples faraway from these landmarks. In this way, the heterogeneous data can by adaptively utilized for target tasks, and, as a result, the generalization ability of transferred detectors is improved. Fig. 16 gives some quantitative results of the auxiliary detector and the target one, respectively.

D. Experiments on ETHZ for Scene Adaptiveness

Here, we experimentally validate our system on two test sequences for busy shopping streets in Fig. 15. We compare our approach with a naïve approach, i.e., replacing CovBoost with AdaBoost in Algorithm 3, but training AdaBoost with the auxiliary positive and the target negative data. The auxiliary detector is considered as a fair baseline to compare the ability of transferred detectors.

Fig. 15(a) shows performance plots. Before adaptation, the power of the auxiliary detector is very limited, as its score is

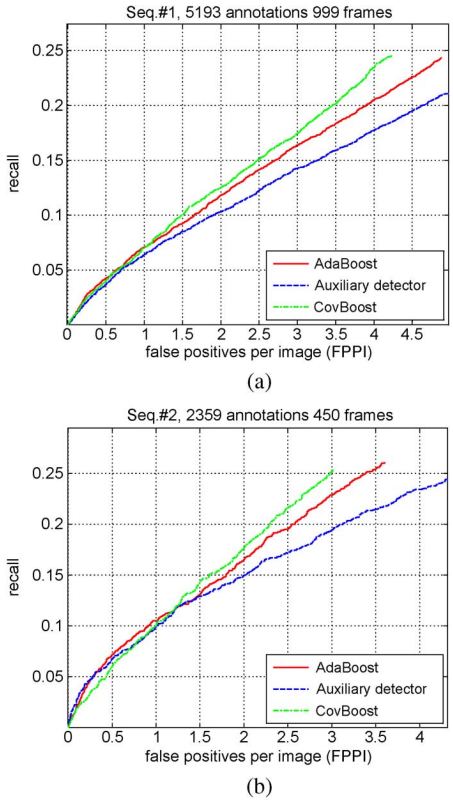


Fig. 15. Performance of different approach for scene adaptiveness task and baseline for the test sequence.

not distinctive enough to separate some human-like shapes, i.e., dust bins and rectangle-like windows. As for the adapted detectors, slightly better result is obtained by the detectors trained by AdaBoost. CovBoost achieves a even better result from using the auxiliary negative examples than AdaBoost on target data only. The plots for sequence #2 also corroborates the advantage of our approach.

Fig. 17 shows two sequences with street scenes in ETHZ dataset. As for the AdaBoost based method, some types of objects tend to be misclassified. For instance, most of FPs in Fig. 17(a) are the trunks. The low contrast feature is the main challenge of Fig. 17(b). Although there are some FPs in Fig. 17(d), our method still achieves less FPs than AdaBoost based method.

Fig. 16. Sample results on *Dataset S7 view3* in PETS2007. (a) Auxiliary detector. (b) Transferred detector.

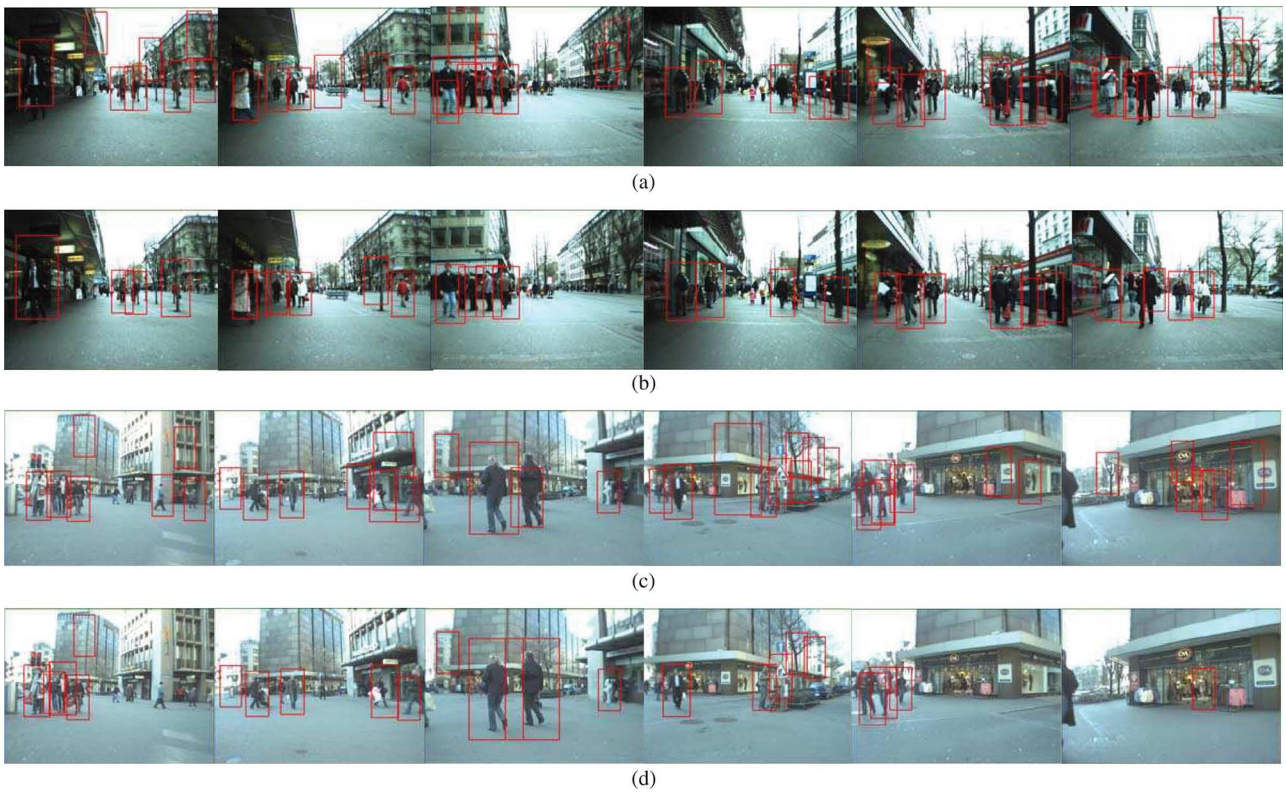


Fig. 17. Scene adaptiveness evaluated on ETHZ dataset under different scenes. (a), (b) Comparative detection results on seq.#1. (c), (d) Detection results on seq.#2 is compared. For more details, please refer to Section V-D.

E. Extension to Other Objects

To demonstrate the generalization ability, we apply the proposed approach to face detection with view-angle change. The 100-profile face (with reflection) collected from web are considered as the target data; while the frontal face examples in [37] are used as the auxiliary data. We use the Haar features [37], because these features show the good ability in face detection. The experiment is carried on CMU profile dataset. It should be noted that this dataset contains a range of view angles of faces, and most of baseline systems [19], [31] divide these faces into subclasses to reduce variations.

The appearance of some profile faces is totally different from the frontal faces; thus, this is not essentially an adaptiveness problem, but a new visual task. The curve of the frontal face detector was not plotted in Fig. 18, because the frontal face detector almost rejects all image patches as negative examples. The improvement of result (10% at 100 false positives) illustrates the benefit of using auxiliary data: even by using a small number of profile face images in training, a portion of profile faces can be detected correctly. Since our profile face detector is handled in a naïve approach (it does not divide the nonfrontal faces into several subclass [19], [31]), our result is natural worse than the baseline [31]. However, the improvement of performance validates the generalization ability to other visual objects.

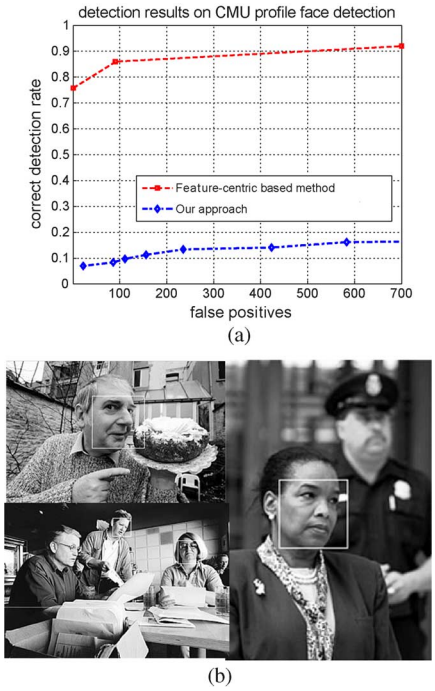


Fig. 18. Performance on CMU profile face detection. (a) Comparison with the baseline result [31]. (b) Some sample results of face detection from our approach.

VI. Conclusion and Future Work

In practice, there exist both necessity and feasibility for transferring a generic detector to new scenarios. The difficulty in collecting universal training data makes a well-trained detector easily fail in a specific scenario. Handling more generic scenes means increasing the complexity of detectors. On the other hand, the possibility to transfer a generic detector also comes from two facts: 1) instances in new scenarios may still share local patches with auxiliary data and 2) specific application scenes have limited backgrounds.

To exploit the possibility, this paper investigates how to transfer boosting-style detectors specifically across viewpoints and scenes. The underlying truth is that the weak classifier corresponds to a local image patch, which is assumed to be shared across viewpoints. By formulating the adaptiveness as covariate shift problem, we propose CovBoost to transfer the auxiliary data in updating detectors. The effectiveness of the proposed method is evaluated on two types of datasets, synthesis and real datasets, from two aspects, i.e., the intuitive efficacy of CovBoost in handling the disparity of data, and the quantitative validation for transferring visual object detectors. The results show that the proposed method can impressively improve the performance of generic detectors across viewpoints and scenes.

It is well known in statistics that importance sampling would be challenged, when the proposal density is far different from the real distribution [24]. A possible solution is to utilize 3-D object model for estimating the underlying mechanism across viewpoints. Therefore, more general transformations of objects should be tested in feature shift.

One of the shortcomings of the proposed method is that, rather than parameterizing the auxiliary knowledge, CovBoost, the data-level transfer learning, requires that auxiliary data should be stored. Moreover, CovBoost does not have the ability of online learning to handle the time-varying scenes. To handle these drawbacks, we are preparing to integrate online updating ability into parameterized CovBoost in our future work.

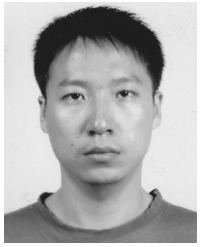
Acknowledgment

The authors would like to thank Dr. Y. Gong for useful discussions. The authors would also like to thank the associate editor and the anonymous reviewers, whose comments helped to improve paper greatly.

References

[1] [Online]. Available: http://pets2007.net
[2] A. Ahmed, K. Yu, W. Xu, Y. Gong, and E. Xing, "Training hierarchal feed-forward visual recognition models using transfer learning from pseudo-tasks," in *Proc. Eur. Conf. Comput. Vision*, 2008, vol. 3, pp. 69–82.
[3] R. Ando and T. Zhang, "A framework for learning predictive structures from multiple tasks and unlabeled data," *J. Mach. Learing Res.*, vol. 6, pp. 1817–1853, 2005.
[4] T. Bakker, "Task clustering and gating for bayesian multitask learning," *J. Mach. Learning Res.*, vol. 4, 2003.
[5] S. Ben-David and R. Schuller, "Exploiting task relatedness for multiple task learning," in *Proc. Annu. Conf. Learning Theor.*, 2003, pp. 567–580.
[6] A. Blum and T. Mitchell, "Combining labeled and unlabeled data with co-training," in *Proc. Annu. Conf. Computational Learning Theor.*, 1998, pp. 92–100.
[7] L. Breiman, "Random forest," *Mach. Learn.*, vol. 45, pp. 5–32, 2001.
[8] R. Caruana, "Multi-task learning," *Mach. Learn.*, vol. 28, no. 1, pp. 41–75, 1997.
[9] W. Dai, Q. Yang, G.-R. Xue, and Y. Yu, "Boosting for transfer learning," in *Proc. Int. Conf. Mach. Learning*, 2007, pp. 193–200.
[10] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, San Diego, CA, 2005, vol. 1, pp. 886–893.
[11] A. Ess, B. Leibe, and L. V. Gool, "Depth and appearance for mobile scene analysis," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2007, vol. 1, pp. 446–453.
[12] M. Everingham, A. Zisserman, C. K. I. Williams, and L. V. Gool, "The PASCAL visual object classes challenge 2006 (VOC2006) results." [Online]. Available: http://www.pascal-network.org/challenges/VOC/voc2006/results.pdf
[13] A. Farhadi and M. Tabrizi, "Learning to recognize activities from the wrong view point," in *Proc. Eur. Conf. Comput. Vision*, 2008, vol. 1, pp. 154–166.
[14] L. Fei-Fei, R. Fergus, and P. Perona, "One-shot learning of object categories," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 28, no. 4, pp. 594–611, Apr. 2006.
[15] Y. Freund and R. E. Schapire, "A short introduction on boosting," *Proc. J. Jpn. Soc. Artif. Intell.*, vol. 14, pp. 771–780, 1999.
[16] Y. Freund and R. E. Schapire, "A decision-theoretic generalization of on-line learning and an application to boosting," in *Proc. Eur. Conf. Computational Learning Theor.*, 1995, vol. 2, pp. 23–37.
[17] H. Grabner and H. Bischof, "On-line boosting and vision," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2006, vol. 1, pp. 17–22.
[18] T. Heskes, "Empirical bayes for learning to learn," in *Proc. Int. Conf. Mach. Learning*, 2000, pp. 367–375.
[19] C. Huang, H. Ai, Y. Li, and S. Lao, "Vector boosting for rotation invariant multi-view face detection," in *Proceedings. 10th IEEE Int. Conf. Comput. Vis.*, Beijing, China, 2005, vol. 1, pp. 446–453.
[20] C. Huang, H. Ai, T. Yamashita, S. Lao, and M. Kawade, "Incremental learning of boosted face detector," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2007, vol. 2, pp. 1–8.
[21] O. Javed, S. Ali, and M. Shah, "Online detection and classification of moving objects using progressively improving detectors," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2005, vol. 1, pp. 696–701.
[22] C. Leistner, H. Grabner, and H. Bischof, "Semi-supervised boosting using visual similarity learning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2008, pp. 1–8.
[23] A. Levin, P. viola, and Y. Freund, "Unsupervised improvement of visual detector using co-training," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2003, vol. 2, pp. 626–633.

[24] D. J. Mackay, *Information Theory, Inference and Learning Algorithms*, 1st ed. Cambridge, U.K.: Cambridge Univ. Press, 2003.
[25] N. Oza and S. Russel, "Online bagging and boosting," *Artif. Intell. Stat.*, 2001.
[26] J. Pang, Q. Huang, and S. Jiang, "Multiple instance boost using graph embedding based decision stump for pedestrian detection," in *Proc. Eur. Conf. Comput. Vision*, 2008, vol. 4, pp. 541–542.
[27] C. Papageorgiou, M. Oren, and T. Poggio, "A general framework for object detection," in *Proc. IEEE Int. Conf. Comput. Vis.*, 1998, vol. 1, pp. 555–562.
[28] G. Perkins, *Transfer of Learning*, 2nd ed. : Int. Encyclopedia of Education, 1992.
[29] L. Rosenstein and Z. Marx, "To transfer or not to transfer," *Tenique Rep.*, 2005.
[30] H. Rowley, S. Baluja, and T. Kanade, "Neural network-based face dtection," *IEEE Pattern Anal. Mach. Intell.*, vol. 20, no. 1, pp. 22–38, Jan. 1998.
[31] H. Schneiderman, "Feature-centric evaluation for efficient cascaded object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2004, pp. 29–36.
[32] M. Sugiyama, M. Krauledat, and K. R. Müller, "Covariate shift adaptation by importance weighted cross validation," *J. Mach. Learning Res.*, vol. 8, pp. 985–1005, 2007.
[33] S. Thrun, Is Learning the n-th Thing Any Easier Than Learning the First? vol. 8, pp. 640–646, 1996.
[34] A. Torralba, K. P. Murphy, and W. T. Freeman, "Sharing features: Efficient boosting procedures for multiclass object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2004, vol. 2, pp. 762–769.
[35] O. Tuzel, F. Porikli, and P. Meer, "Human detection via classification on Riemannian manifolds," in *Proc. IEEE Conf. Comput. Vison Pattern Recognition*, 2007, pp. 1–8.
[36] V. Vapnik, *The Nature of Statistic Learning Theory*, 2nd ed. New York: Springer-Verlag, 1995.
[37] P. Viola and M. Jones, "Rapid object detection using a boosted cascade of simple features," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2001, vol. 1, pp. 511–518.
[38] P. Viola, J. Platt, and C. Zhang, "Multiple instance boosting for object detection," *Neural Inf. Process. Syst.*, pp. 1417–1426, 2005.
[39] A. Wilson, A. Fern, S. Ray, and P. tadepalli, "Multi-task reinforecement learning: A hierarchical bayesian approach," in *Proc. Int. Conf. Mach. Learning*, 2007, vol. 227, pp. 1015–1022.
[40] B. Wu and R. Nevatia, "Improving part based object detection by unsupervised, online boosting," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2007.
[41] C. Zhang, R. Hamid, and Z. Zhang, "Taylor expansion based classifier adaption: Application to person detection," in *Proc. Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, 2008, pp. 1–8.



Junbiao Pang received the B.S. and M.S. degrees in computational fluid dynamics and computer science from the Harbin Institute of Technology, Harbin, China, in 2002 and 2004, respectively. He is currently working toward the Ph.D. degree at the Institute of Computing Technology, Chinese Academy of Sciences, Beijing, China.

His research areas include computer vision, multimedia and machine learning, and he has authored or coauthored approximately ten technical papers.

Mr. Pang is a reviewer of various international journals, including the IEEE TRANSACTIONS ON SYSTEMS, MAN AND CYBERNETICS—PART B and the IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY.

Qingming Huang (SM'08) received the B.S. degree in computer science and Ph.D. degree in computer engineering from Harbin Institute of Technology, Harbin, China, in 1988 and 1994, respectively.

He is currently a Professor with the Graduate University of the Chinese Academy of Sciences (CAS), Beijing, China, and an Adjunct Research Professor with the Institute of Computing Technology, CAS. He has been granted by China National Funds for Distinguished Young Scientists in 2010. He has authored or coauthored more than 170 academic papers in prestigious international journals and conferences. His research areas include multimedia video analysis, video adaptation, image processing, computer vision, and pattern recognition

Dr. Huang is a reviewer for the IEEE TRANSACTIONS ON MULTIMEDIA and the IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY. He has served as a TPC member for well-known conferences, including ACM Multimedia, CVPR, ICCV, and ICME.

Shuicheng Yan (SM'09) is currently an Assistant Professor with the Department of Electrical and Computer Engineering, National University of Singapore, Singapore, and the founding lead of the Learning and Vision Research Group. His research areas include computer vision, multimedia, and machine learning, and he has authored or coauthored approximately 190 technical papers over a wide range of research topics.

Dr. Yan is an associate editor of the IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY and has been serving as the guest editor of the special issues for TMM and CVIU. He was the recipient of Best Paper Awards from ACM MM10, ICME10, and ICIMCS'09 and the PRE-MIA 2008 Best Student Paper Award, the winner prize of the classification task in PASCAL VOC2010, and the honorable mention prize of the detection task in PASCAL VOC2010.

Shuqiang Jiang (SM'08) received the M.S. degree from the College of Information Science and Engineering, Shandong University of Science and Technology, Shandong, China, in 2000, and the Ph.D. degree from the Institute of Computing Technology, Chinese Academy of Sciences, Beijing, China, in 2005.

He is currently a Faculty Member with Digital Media Research Center, Institute of Computing Technology, Chinese Academy of Sciences (CAS), Beijing, China. He is also with the Key Laboratory of Intelligent Information Processing, CAS. His research interests include multimedia processing and semantic understanding and pattern recognition.

Lei Qin (M'06) received the B.S. and M.S. degrees in mathematics from the Dalian University of Technology, Dalian, China, in 1999 and 2002, respectively, and the Ph.D. degree in computer science from the Institute of Computing Technology, Chinese Academy of Sciences, Beijing, China, in 2008.

He is currently a Faculty Member with the Key Laboratory of Intelligent Information Processing, Institute of Computing Technology, Chinese Academy of Sciences, Beijing, China. His research interests include image/video processing, computer vision, and pattern recognition. He has authored or coauthored over ten technical papers in the area of computer vision.