

# Unsupervised Selective Transfer Learning for Object Recognition

Wei-Shi Zheng, Shaogang Gong and Tao Xiang

School of Electronic Engineering and Computer Science  
Queen Mary University of London, London E1 4NS, UK  
{jason,sgg,txiang}@dcs.qmul.ac.uk

**Abstract.** We propose a novel unsupervised transfer learning framework that utilises unlabelled auxiliary data to quantify and select the most relevant transferrable knowledge for recognising a target object class from the background given very limited training target samples. Unlike existing transfer learning techniques, our method does not assume that auxiliary data are labelled, nor the relationships between target and auxiliary classes are known *a priori*. Our unsupervised transfer learning is formulated by a novel kernel adaptation transfer (KAT) learning framework, which aims to (a) extract general knowledge about how more structured objects are visually distinctive from cluttered background regardless object class, and (b) more importantly, perform selective transfer of knowledge extracted from the auxiliary data to minimise negative knowledge transfer suffered by existing methods. The effectiveness and efficiency of the proposed approach is demonstrated by performing one-class object recognition (object vs. background) task using the Caltech256 dataset.

## 1 Introduction

Object recognition in unconstrained environments is a hard problem largely due to the vast intra-class diversities in object forms and appearance. Consequently, to learn a classifier for recognition, one typically needs hundreds or even thousands of training samples in order to account for the visual variability of objects within each class [1]. However, labelling large number of training samples is not only expensive but also not always viable because of difficulties in obtaining data samples for rare object classes. Inspired by human’s ability to learn new object categories with very limited samples [2], recent studies have started to focus on transfer learning [3–8], with which a model is designed to transfer object class knowledge from previously learned models to newly observed images of either the same object class or different classes. The overall aim of transfer learning is to maximise available information about an object class given very sparse training samples by utilising shared and relevant knowledge from other object classes and/or the same class of significant intra-class appearance variation. Two non-trivial challenges surface: how to quantify and compute ‘shared relevant’ inter- and intra-class object appearance knowledge, and under what assumptions.

In this paper, we consider the object transfer learning for the extensively studied one-class recognition [9, 7, 3], which learns a model that is able to detect

different and unknown object categories (target categories) against cluttered background. We assume that only very limited target training samples but a number of unlabelled images of other non-background object categories (auxiliary categories) are available, where the “unlabelled” means the exact labels of auxiliary data and the relationships such as the hierarchical category structure between the auxiliary and target categories are unknown.

The proposed model addresses two significant limitations of existing transfer learning methods. First, most existing methods require that auxiliary classes are labelled and the relationships between target and auxiliary classes are known *a priori*, e.g. cross-domain but from the same categories [10–13] or cross-category but relevant using hierarchy category structure [9] (e.g. detecting giraffes using other four-leg animals as auxiliary data). However, identifying the relationships between a target category and auxiliary categories is non-trivial and not always possible – relevant object categories with shared knowledge to the target class may not be available in the auxiliary data. Although a few recent studies demonstrate that this assumption can be relaxed [7, 3], these methods are fully supervised so the problem is somewhat averted rather than solved as they all require a costly exhaustive labelling of *all* the auxiliary data to mitigate unknown relationships. However, exhaustive labelling is not always available nor viable. Second, no measures are taken by existing transfer learning techniques to avoid negative knowledge transfer [14, 13], which reduces rather than enhances the performance of an object recogniser learned from target class samples alone. Clearly there is a need for quantifying the usefulness of auxiliary knowledge and therefore explicitly selecting the relevant one in order to significantly alleviate negative transfer.

To overcome these limitations of existing models, we propose an unsupervised transfer learning model capable of 1) identifying automatically transferrable knowledge to be extracted if both class labels of the auxiliary data and any relationships between the auxiliary and target object classes are unknown, and 2) more important, quantifying and selecting the most effectively relevant auxiliary knowledge (not all are relevant) and combining them with target specific knowledge learned from very limited training samples in constructing optimal target object recognisers. To that end, our unsupervised transfer learning is formulated in a novel Kernel Adaptation Transfer (KAT) learning framework. For the first objective, we exploit an observation that regardless how dissimilar the auxiliary object categories may be from a target category, there is shared general knowledge (structural characteristics) of objects, which are visually distinctive from less structured and more random background clutters. To extract such general knowledge from unlabelled auxiliary data, we perform clustering on the auxiliary data and the general knowledge is computed as an ensemble of local auxiliary knowledge that distinguishes objects from each cluster from cluttered background. For the second objective, we represent the extracted auxiliary knowledge using multiple kernel functions, and combine them with any target specific knowledge represented by a kernel function learned directly from a handful of target class samples. Critically, since the auxiliary data classes

may be ineffective to the target class which we have no knowledge of *a priori*, the usefulness of all auxiliary kernel functions is evaluated automatically and only relevant kernels are selected by our transfer kernel adaptation framework, in which a hypothesis constraint between the weight of target kernel and the weights of auxiliary kernels and a partial sparsity constraint are introduced in order to minimise negative knowledge transfer.

In summary, the proposed unsupervised kernel adaptation transfer learning model makes two main contributions: (1) it provides a novel way of extracting general auxiliary knowledge from unlabelled auxiliary categories; (2) it significantly alleviates negative transfer due to its ability of quantifying and selecting transferrable knowledge. Our results on Caltech256 dataset [15] demonstrate that: (a) Our method is able to extract selectively useful transferrable knowledge from unlabelled auxiliary data to improve the recognition performance of each individual target object model trained by very limited samples. The selection is different for different target classes. (b) Our unsupervised transfer learning model outperforms existing unsupervised transfer learning model in terms of recognition performance, robustness to negative transfer and computational cost; and is comparable to or better than supervised transfer learning with fully labelled auxiliary data.

## 2 Related Work

Most existing transfer learning methods for object recognition are supervised (i.e. they require labelled auxiliary data), and make the assumption that the auxiliary classes must be related to each target object class. In [9], hierarchical category structure is needed for selecting auxiliary classes. Similarly, Stark et al. [6] transfer the variance of object shape from closely related object classes using a generative model inspired by Constellation model [16]. Alternatively, in [4, 5] a model-free approach is taken for which common attribute information has to be specified manually. In comparison, our KAT model does not assume any relationship between auxiliary object classes and a target class, and the knowledge learned and transferred by our method can be extracted from auxiliary object classes unrelated (i.e. not similar) to the target class.

There have been a number of attempts to transfer general knowledge about object visual appearance using unrelated object categories. Fei-Fei et al. [7] formulated an one-shot Bayesian learning framework based on Constellation models. Models are learned from labelled auxiliary categories and the model parameters are averaged and used in the form of a prior probability density function in the parameter space for learning a target object class. As pointed out by Bart and Ullman [3], since only one prior distribution is used, this approach can be biased by the dominant or common auxiliary categories, and the method is unable to detect and prevent this from happening. The feature adaptation method proposed by Bart and Ullman [3] overcomes this problem by learning a set of discriminative features from each labelled auxiliary category and selecting the features extracted from target category samples towards those discriminative auxiliary features. The adaptation is meaningful even if only one of 100 auxiliary

categories is related to the target class. Similar to [3], our method can avoid the problem of bias towards dominant auxiliary categories. The difference is that we do not need to assume the availability of any related auxiliary categories since the knowledge transferred by our method is concerned with how objects in general can be distinguished from background. Moreover, no labelling is required for auxiliary data using our model.

The most closely related work is the self-taught learning (STL) method proposed by Raina et al. [17] which also does not require labelling of auxiliary data, even though another unsupervised transfer learning method was also proposed recently [18]. The method in [18] is transductive and requires testing data to be involved in the training stage which may be considered less desirable. As compared to STL, our KAT model is superior in three ways. (1) The knowledge extracted by STL is a set of feature bases that best describe how objects look alike. In contrast our KAT model extracts the general knowledge about how objects are distinguishable from cluttered background. It is thus more suitable for the one-class object recognition task. This is supported by our experimental results shown in Section 5. (2) Most importantly, the feature bases learned by STL from auxiliary data are trusted blindly and used for target category sample representation without discrimination. In contrast, our kernel adaptation method provides a principled way of quantifying and selecting the most useful auxiliary information relevant to different target categories for discrimination. As a result, our KAT model is more robust against negative knowledge transfer. (3) The STL model is computationally much more expensive than our KAT model, especially on high-dimensional features, as  $l_1$ -norm based optimisation is required.

Our kernel adaptation transfer (KAT) is a multiple kernel learning (MKL) method. KAT differs from the non-transfer MKL [19, 20] in that it is specially designed with two proposed constraints for unsupervised cross-category transfer learning to fuse a target kernel with a number of auxiliary kernels built based on lots of auxiliary data that may be irrelevant to target category. Though several work on kernel function learning has been exploited before for transfer learning [21, 13, 10, 11], KAT is formulated for unsupervised cross-category transfer learning, whilst [21, 13, 10, 11] assume that the target class data and auxiliary data are from different domains of the same category (e.g. news video footage from different countries) and they can neither be applied directly to nor easily extended for our cross-category unsupervised transfer learning task.

### 3 Kernel Adaptation for Knowledge Transfer

Let us first describe in this section our general kernel adaptation transfer learning framework before formulating a specific model for an object recognition problem in the next section. Given a training dataset  $\{\mathbf{x}_i^t, y_i^t\}_{i=1}^N$  where  $y_i^t \in \{1, -1\}$  is the label of  $\mathbf{x}_i^t$  and we call  $y_i^t = -1$  the negative class (e.g. background) and  $y_i^t = 1$  the target class (e.g. object). We wish to transfer knowledge extracted from an unlabelled (and possibly also irrelevant) auxiliary dataset  $\{\mathbf{x}_j^a\}$  to the target class data, where  $\{\mathbf{x}_j^a\}$  do not contain any samples of the target class. We consider that the auxiliary knowledge is represented by a set of auxiliary

mappings  $f_s, s = 1, \dots, m$  which are learned from auxiliary data such that  $f_s(\mathbf{x}_i^t), s = 1, \dots, m$  are the candidate transferrable knowledge for each training target data  $\mathbf{x}_i^t$ . The form of the auxiliary mapping  $f_s(\mathbf{x}_i^t)$  depends on the specific transfer learning problem, for example a feature vector. Given any auxiliary mappings extracted from unlabelled auxiliary data, our objective is to quantify and select the transferrable ones and combine them with the data  $\{\mathbf{x}_i^t, y_i^t\}_{i=1}^N$  in order to get better recognition performance. More specifically, the problem becomes learning a mapping  $g$  by combining  $f_s(\mathbf{x}_i^t)$  and  $\mathbf{x}_i^t$  in order to generate a new  $d$ -dimensional vector as follows:

$$g : (\mathbf{x}_i^t, \{f_s(\mathbf{x}_i^t)\}_{s=1}^m) \longrightarrow \mathbf{z}_i^t \in \mathfrak{R}^d. \quad (1)$$

In this paper, we learn such a  $g$  via kernel methods. It is motivated because (a) rather than being explicitly defined, by formulating a kernel framework,  $g$  can be implicitly induced by a Mercer kernel function, which we call the transfer kernel function in this paper; (b) by optimising the transfer kernel, we obtain a principled way for quantifying and selecting transferrable knowledge, which is critical for minimising negative knowledge transfer.

We now show why and how a unsupervised transfer learning problem can be formulated as a multi-kernel adaptation problem. From the statistical point of view, by computing a Mercer kernel matrix  $\mathbf{K} = (\kappa(\mathbf{x}_i^t, \mathbf{x}_j^t))_{ij}$ , each entry  $\kappa(\mathbf{x}_i^t, \mathbf{x}_j^t)$  can be considered to be proportional to the underlying pairwise distribution of any two data points  $\mathbf{x}_i^t, \mathbf{x}_j^t$  in the training dataset, that is,

$$p(\mathbf{x}_i^t, \mathbf{x}_j^t) = C \cdot \kappa(\mathbf{x}_i^t, \mathbf{x}_j^t), \quad (2)$$

where  $C$  is some distribution normalizer. Assuming for generating  $p(\mathbf{x}_i^t, \mathbf{x}_j^t)$  there is a latent function variable  $f$  with probability density distribution  $p(f)$ , the distribution  $p(\mathbf{x}_i^t, \mathbf{x}_j^t)$  can then be expressed as:

$$\begin{aligned} p(\mathbf{x}_i^t, \mathbf{x}_j^t) &= \oint_f [p(\mathbf{x}_i^t, \mathbf{x}_j^t, f)] = \oint_f [p(\mathbf{x}_i^t, \mathbf{x}_j^t | f) p(f)] \\ &= \oint_f [\tilde{p}_f(f(\mathbf{x}_i^t), f(\mathbf{x}_j^t)) p(f)] \approx \sum_{f_s, s=0, \dots, m} \tilde{p}_{f_s}(f_s(\mathbf{x}_i^t), f_s(\mathbf{x}_j^t)) P(f_s), \end{aligned} \quad (3)$$

where we define  $\tilde{p}_f(f(\mathbf{x}_i^t), f(\mathbf{x}_j^t)) = p(\mathbf{x}_i^t, \mathbf{x}_j^t | f)$  being a conditional density function with function  $f$  imposed on data,  $f_0$  is the identity function (i.e.  $f_0(\mathbf{x}) = \mathbf{x}$ ) and  $f_s, s = 1, \dots, m$  are the auxiliary mappings. The discrete approximation in Eqn. (3) is based on the assumption that the knowledge ( $f_s(\mathbf{x}_i^t)$ ) extracted from target and auxiliary data can be used to infer the underlying density function  $p(\mathbf{x}_i^t, \mathbf{x}_j^t)$ . Note that since  $f_0$  is the identity function, we allow  $\tilde{p}_{f_0}(\mathbf{x}_i^t, \mathbf{x}_j^t)$  to be different from  $p(\mathbf{x}_i^t, \mathbf{x}_j^t)$ . This is because  $\tilde{p}_{f_0}(\mathbf{x}_i^t, \mathbf{x}_j^t)$  is an approximation of  $p(\mathbf{x}_i^t, \mathbf{x}_j^t)$ , which is the unknown intrinsic density function of target object class.

The above model provides a way to estimate the intrinsic density function by transferring auxiliary knowledge for target class in the form of multi-kernel adaptation. Specifically, we let

$$\tilde{p}_{f_s}(f_s(\mathbf{x}_i^t), f_s(\mathbf{x}_j^t)) = C_s \cdot \kappa_s(f_s(\mathbf{x}_i^t), f_s(\mathbf{x}_j^t)), \quad s = 0, \dots, m \quad (4)$$

for some distribution normalizer  $C_s$  with respect to kernel  $\kappa_s$ . Let  $b_s = C^{-1} \cdot C_s \cdot P(f_s), s = 0, \dots, m$ , then we can replace  $p(\mathbf{x}_i^t, \mathbf{x}_j^t)$  in Eqn. (3) using Eqn. (2)

and  $\tilde{p}_{f_s}(f_s(\mathbf{x}_i^t), f_s(\mathbf{x}_j^t))$  using Eqn. (4). Eqn. (3) can thus be rewritten as:

$$\kappa(\mathcal{A}_i^t, \mathcal{A}_j^t) = b_0 \cdot \kappa_0(\mathbf{x}_i^t, \mathbf{x}_j^t) + \sum_{s=1}^m b_s \cdot \kappa_s(f_s(\mathbf{x}_i^t), f_s(\mathbf{x}_j^t)), \quad (5)$$

where we denote  $\kappa(\mathcal{A}_i^t, \mathcal{A}_j^t) = \kappa(\mathbf{x}_i^t, \mathbf{x}_j^t)$ ,  $\mathcal{A}_i^t = \{\mathbf{x}_i^t, \{f_s(\mathbf{x}_i^t)\}_{s=1}^m\}$ ,  $\kappa_0$  is the kernel function on data  $\mathbf{x}_i^t$  itself, and  $\kappa_s$  is constructed using the transferrable knowledge  $f_s(\mathbf{x}_i^t)$  extracted from the auxiliary data. We call  $\kappa$  the kernel transfer function.

Learning optimal non-negative weights  $\{b_i\}_{s=0}^m$  from data is equivalent to learning an optimal transfer kernel matrix  $\mathbf{K} = (\kappa(\mathcal{A}_i^t, \mathcal{A}_j^t))_{i,j}$  via the combination of kernel matrices  $\mathbf{K}_s = (\kappa_s(f_s(\mathbf{x}_i^t), f_s(\mathbf{x}_j^t)))_{i,j}$ ,  $s = 0, \dots, m$ . By exploiting the close relationship between kernel product and the probability of pairwise data, we learn an optimal transfer kernel matrix such that the entry value of intra-class pairs in the kernel matrix is as large as possible while those of the other entries are as small as possible. Let  $\mathbf{b} = (b_0, b_1, \dots, b_m)^T$ , we thus have:

$$\mathbf{b} = \arg \max_{\mathbf{b}} \frac{\mathbf{K}(\cdot)^T \mathbf{K}_{opt}^+(\cdot)}{\mathbf{K}(\cdot)^T \mathbf{K}_{opt}^-(\cdot) + \alpha \cdot \mathbf{b}^T \mathbf{b}}, \quad (6)$$

*s.t.*  $b_s \geq 0$ ,  $s = 0, \dots, m$ ,  $\alpha > 0$ ,

where  $\mathbf{K}(\cdot)$  represents a vectorised kernel matrix and  $\mathbf{K}(\cdot) = \Psi \mathbf{b}$ ,  $\Psi = [\mathbf{K}_0(\cdot), \mathbf{K}_1(\cdot), \dots, \mathbf{K}_m(\cdot)]$ , and  $\mathbf{K}_{opt}^+$  and  $\mathbf{K}_{opt}^-$  are defined as follows:

$$\mathbf{K}_{opt}^+(i, j) = \begin{cases} 1 & y_i^t = y_j^t, \\ 0 & y_i^t \neq y_j^t, \end{cases} \quad \mathbf{K}_{opt}^-(i, j) = \begin{cases} 0 & y_i^t = y_j^t, \\ 1 & y_i^t \neq y_j^t. \end{cases} \quad (7)$$

Note that the regularisation term  $\alpha \cdot \mathbf{b}^T \mathbf{b}$  is necessary in order to avoid learning a trivial  $\mathbf{b}$  that one entry of  $\mathbf{b}$  is always 1 and the others are zero. In this work,  $\alpha$  is set to 0.01.

Rather than directly learning  $b$  based on the above multiple kernel learning criterion, we further introduce two additional constraints to address an imbalanced kernel fusion problem as follows. Different from most non-transfer MKL work [19, 20], our transfer learning is to combine a single target kernel ( $\kappa_0$ ) and a large amount of auxiliary kernels ( $\kappa_s$ ,  $s \geq 1$ ). Hence ineffective/harmful auxiliary kernels could have large effect on this kind of fusion, which would result in negative transfer. In order to balance the effect of the auxiliary kernels on the combined kernel, we introduce two constraints are described below:

1. **Hypothesis Constraint.** It constrains the order between the weight of target kernel and the weights of auxiliary kernels as follows:

$$b_0 \geq b_s, \quad s \geq 1, \quad (8)$$

which enforces that more weight is given to the kernel function built directly from the target class than any other auxiliary kernels. This is intuitive as one wants to trust more on the limited target class data than any auxiliary mapping when auxiliary data are unlabeled.

2. **Partial Sparsity Constraint.** As a second constraint, which is the most important, we impose the following partial  $l_1$ -norm based sparsity penalty on  $\mathbf{b}$  on auxiliary data for minimization:

$$\mathbf{b}^T \mathbf{1}_0, \quad \mathbf{1}_0 = [0, 1, 1, \dots, 1]^T. \quad (9)$$

Note that this sparsity penalty is partial because it does not apply to the target class kernel. The objective of this partial sparsity constraint is to allow only a small portion of the relevant auxiliary kernels to be combined with the target class kernel. In our experiments, we demonstrate that these two constraints play a critical role in minimising negative knowledge transfer.

With those two constraints, our proposed *kernel adaptation transfer* (KAT) model is formulated as:

$$\mathbf{b} = \arg \max_{\mathbf{b}} \frac{\mathbf{K}(\cdot)^T \mathbf{K}_{opt}^+(\cdot)}{\mathbf{K}(\cdot)^T \mathbf{K}_{opt}^-(\cdot) + \alpha \cdot \mathbf{b}^T \mathbf{b} + \lambda \cdot \mathbf{b}^T \mathbf{1}_0}, \quad (10)$$

$$s.t. \ b_0 \geq b_s \geq 0, \ s = 1, \dots, m, \ \alpha > 0, \ \lambda \geq 0.$$

Solving the above optimisation problem is nontrivial. However, by reformulating Criterion (10) alternatively as follows, a quadratic programming solver [22] can be exploited to find the solution:

$$\mathbf{b} = \arg \min_{\mathbf{b}} \mathbf{b}^T \Psi^T \mathbf{K}_{opt}^-(\cdot) + \alpha \cdot \mathbf{b}^T \mathbf{b} + \lambda \cdot \mathbf{b}^T \mathbf{1}_0, \quad (11)$$

$$s.t. \ \mathbf{b}^T \Psi^T \mathbf{K}_{opt}^+(\cdot) = 1; \ b_0 \geq b_s \geq 0, \ s = 1, \dots, m, \ \alpha > 0, \ \lambda \geq 0.$$

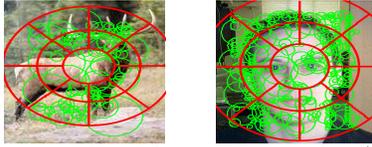
where  $\lambda$  is a free parameter that controls the strength of the partial sparsity penalty (thus how much auxiliary knowledge can be transferred). Specifically the smaller  $\lambda$  is, the more auxiliary knowledge can be transferred to target data, and no auxiliary knowledge can be transferred if  $\lambda = +\infty$ , which would set all weights  $b_s, s = 1, \dots, m$  on auxiliary kernels  $\kappa_s$  to zero.

Finally, we note that except the two proposed constraints for transfer learning, KAT is closely related to the widely adopted kernel alignment method [23], which assumes the optimal entry values of kernel matrix  $\mathbf{K}$  should be 1 for intra-class pairs and 0 otherwise. In comparison, Eqn (10) relaxes this assumption and maximises the ratio between those two types of values. Due to the nature of data distribution (e.g. Caltech 256), the strict assumption made by kernel alignment would not be held and could result in learning a combined kernel being dominated by the large amount of ineffective auxiliary kernels.

## 4 Knowledge Transfer for Object Recognition

We now re-formulate the unsupervised transfer learning framework described in Section 3 into a practical model for the one-class object recognition problem. For such a problem, it is assumed that an image contains either one object from a target class or only background clutter [24, 7, 9]. By applying our transfer learning framework we assume that for each target object category of interest, there are only very limited training samples (from only one sample to no more than a dozen at best) and an auxiliary dataset containing a large number of different object categories, where the auxiliary images are without any class labels and could be irrelevant to the target class. In addition we have a random set of cluttered background images which can be obtained by simply searching “Things” on Google Image [15].

For object representation, we first detect salient feature points using the Kadir and Brady detector [25]. A fixed number of feature points are then retained



**Fig. 1.** The polar structure used for object representation ( $o = 8$ ,  $r = 3$  and 17 bins).

by thresholding the saliency value. The spatial distribution of those points is represented by a polar geometry structure with  $o$  orientational bins and  $r$  radial bins as illustrated in Figure 1. This polar structure is automatically expanded from the center of the detected feature points, i.e.  $(B^{-1} \cdot \sum_{i=1}^B Z_x^i, B^{-1} \cdot \sum_{i=1}^B Z_y^i)$ , where  $B$  is the number of feature points and  $(Z_x^i, Z_y^i)$  is the coordinate of the  $i^{th}$  feature point. The maximum value of the radius is set to  $d_{max} = \max_i \|Z_x^i - B^{-1} \cdot \sum_{i=1}^B Z_x^i\|_2 + \|Z_y^i - B^{-1} \cdot \sum_{i=1}^B Z_y^i\|_2$  so that all feature points are covered in the polar structure. For representing the appearance of the feature points, SIFT descriptors [26] are computed at each feature point, and for each bin in the polar structure a normalized histogram vector is constructed using a codebook with 200 codewords obtained using  $k$ -means. This gives us a 3400 dimensional feature vector ( $\mathbf{x}_i^t$  in Eqn. (5)). Note that this polar object representation scheme is not only much simpler and easier to compute than the Constellation model based representation [16, 24, 27], but more importantly, it is more suitable for our discriminative based kernel adaptation transfer model.

Given the training target and background data  $\{\mathbf{x}_i^t, \mathbf{y}_i^t\}_{i=1}^N$ , we now describe how the auxiliary mappings  $f_s(\mathbf{x}_i^t)$  (see Eqn. (1)) are obtained. We formulate the auxiliary mappings that capture knowledge about how objects look different and distinguishable from any background. We thus formulate the mappings as decision boundaries between groups of objects and cluttered background learned using a Support Vector Machine (SVM). More specifically, the unlabelled auxiliary data is clustered using an ensemble of  $k$ -means, resulting in  $m$  clusters in total. That is, multiple  $k$ -means with different  $k$  values. In this way, the same auxiliary image can belong to different clusters simultaneously. This is intended to reflect the fact that an object can be distinguished from background in many different ways (e.g. colour, texture, shape). For each cluster, denoted as  $\mathcal{G}_s$ , a decision function  $h_s(x)$  is learned by linear SVM. These local (cluster specific) decision functions are then used as the mapping functions, i.e.  $f_s(\mathbf{x}_i^t) = h_s(\mathbf{x}_i^t)$ . Note that outputs of SVM classifier as features are also adopted in [9, 28, 20], but the objectives of using these features are different here.

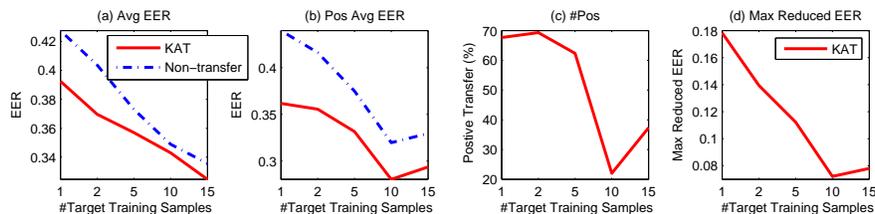
Now with the  $m$  auxiliary mappings  $f_s(\mathbf{x}_i^t)$ , we can quantify them and compute the kernel function  $\kappa(\mathcal{A}_i^t, \mathcal{A}_j^t)$  in Eqn. (5) for the target category, which is a combination of multiple kernels, where we use RBF as the basis kernels. Finally, this kernel function is used to train a SVM for classifying the target object category against cluttered background.

## 5 Experiments

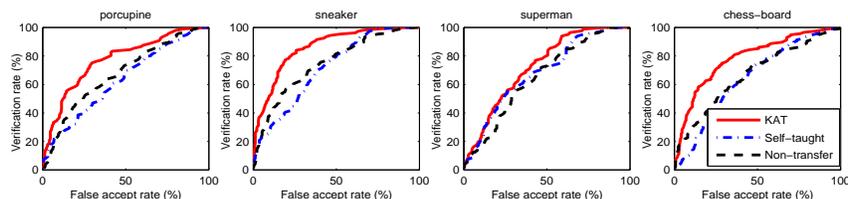
Experiments were conducted on Caltech256 dataset [15]. It contains a broad ranges of objects from 256 object categories and a cluttered background set.

In each experiment, we randomly selected 30 categories as target classes, and selected 10 images from each of the remaining 226 categories to form the auxiliary dataset and their labels were ignored in all the unsupervised transfer learning part of the experiments. This was repeated for 10 times with a different 30 target classes randomly drawn each time, giving 300 one-class object recognition tasks in total. For each target class, unless otherwise stated (as in Figure 2), 5 target images were randomly selected for training and the rest (always more than 80 images) were used for testing. The background set was randomly divided into 3 subsets: 20% of the total number of background images were used for learning the auxiliary kernels, 30% were used for learning the transfer kernel, and the rest 50% were for testing. For different target classes, the same training and testing background image sets were used. The recognition performance was measured using both equal error rate (EER) and receiver-operating characteristic (ROC) curve [16]. For object representation, we selected 300 feature points in each image according to their entropy saliency values [25]. In KAT, for clustering the auxiliary data using an ensemble of  $k$ -means, the value of  $k$  was set to  $\{1, 5, 11, 23, 28, 45, 75\}$  giving a total of 188 clusters. For  $\lambda$  in Eqn. (11), cross validation (CV) was performed when the number of training samples for target object ( $p$ ) is larger than 1. Specifically, two-fold CV was used for  $p = 2$  and three-fold CV was used for  $p = 5, 10, 15$ . Note that the training background images were also used for cross-validation. Then, when the target class training set contains a single sample (i.e. one shot learning),  $\lambda$  was set to 0 (i.e. all auxiliary knowledge is transferred) as cross validation becomes impossible on target object data.

**Kernel adaptation transfer vs. without transfer.** We first compared the performance of our KAT model, learned using both target samples (up to 15) and the unlabelled auxiliary data, against a non-transfer model, namely a SVM classifier trained using only target samples. In this particular comparison, we varied the number of training target category samples from 1 to 15 in order to understand the effect of transfer learning when increasing number of target samples becomes available. The same image descriptor was used for both methods so the performance difference was solely due to using transferred knowledge. The results are shown in Figure 2, where “Avg EER” is the average EER over 300 trials, “Pos Avg EER” is the average EER over those trials when EER was reduced by positive transfer, “#Pos” is the number of trials with positive transfer, and “#Neg” is the number of trials with negative transfer. Figure 2(a) shows that on average the EER was reduced from 0.428 to 0.392 when only one target sample was used from each target class, representing a 8.4% improvement in recognition performance. However, when the number of available target samples increases, the contribution of transfer learning diminishes, shown by the narrowed EER gap between the KAT and the non-transfer model. This is expected as transferring auxiliary kernels has its greatest benefit when the available target training sample is minimal, e.g. with only 1 target sample. Figure 2(c) shows that positive transfer was achieved for 67.7% of the 300 trials for one-shot learning, and reduced to 37.3% when the target training samples was increased to 15



**Fig. 2.** KAT vs Non-transfer model. From left to right, the Avg EER, Pos Avg EER, #Pos (see text for explanation), and Maximum Reduced EER are shown.



**Fig. 3.** KAT model ROC performance curves from four example target classes.

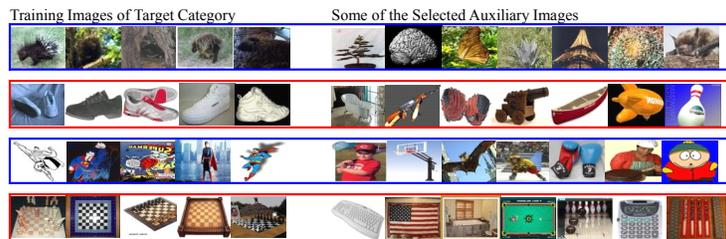
for each category. Interestingly although both the maximum EER reduction per class (Figure 2(d)) and the average EER decrease as the number of data samples increases, the average reduction for the classes with positive transfer was about the same. This suggests that the decrease in the number of positive transfer classes (Figure 2(c)) was the reason for the narrowed EER gap in Figure 2(a).

Figure 3 shows the performance of our KAT on 4 target categories using only 5 target samples for training each. The randomly selected training samples can be seen in the left column of Figure 4 which clearly show that there are huge variations in appearance and view angle for each class. Despite these variations, Figure 3 shows that very good performance was obtained with significant improvement over the non-transfer model. Figure 4 gives some insight into what auxiliary categories have been extracted and transferred to the target classes by the KAT model. In particular, it can be seen that KAT can automatically select object images from related classes if they are present in the auxiliary data (e.g. a butterfly for superman that flies), or from objects that share some similarity in shape or appearance (e.g. bonsai for porcupine, boat for shoe, pool table for chess board) for extracting transferrable knowledge.

**Kernel adaptation transfer vs. self-taught learning.** We compared our KAT model with the most related unsupervised transfer learning method we are aware of, the Self-Taught Learning (STL) model [17]. For STL, we followed

Method	Avg EER	Pos Avg EER	#Pos	#Neg
KAT	0.357	0.332 (0.375)	187	84
STL	0.409	0.388 (0.426)	87	196
Aux	0.388	0.369 (0.419)	123	161
Non-Transfer	0.373	N/A	N/A	N/A

**Table 1.** Comparing KAT with STL [17] and a generic non-background classifier (Aux) given 5 samples per class. The numbers in brackets are the performance of non-transfer model for the target classes for which positive transfer was achieved.



**Fig. 4.** Examples of target categories (left) and the auxiliary data that provide the most transferrable knowledge (right) by KAT. The four target categories from top to bottom are porcupine, sneaker, superman, and chess-board respectively.

the PCA dimension reduction procedure in [17] on the high-dimensional descriptors and exactly the same procedure in [17] to train the self-taught learner. The number of sparse feature bases was set to be the same as the reduced dimensionality and the sparse weights were tuned in  $\{0.005, 0.05, 0.1, 0.5\}$  [17] due to the computational cost of the STL model, as discussed later. As STL is computationally expensive, it is not possible to select the parameter using cross-validation in practice. Instead, we tuned the parameter of STL using the test data to illustrate the best performance STL can possibly achieve. In other words, lower performance is expected for STL if cross validation is used. The performances of the two unsupervised transfer learning methods are compared in Table 1. It shows that a majority (196) of the 300 object recognition tasks result in negative transfer using STL. As a result, the overall performance of STL was actually worse than the baseline non-transfer model (0.409 vs. 0.373). This is because STL cannot control how much auxiliary knowledge should be transferred for different target categories. In comparison, because of measuring the usefulness of and automatically selecting transferrable knowledge for different target classes, our KAT achieved far superior results with only 84 out of 300 tasks lead to negative transfer. Table 1 also shows that even for those positive transfer tasks, KAT yields higher improvement over STL (0.043 vs. 0.038 in EER reduction). This suggests that the knowledge transferred by our method, which aims to discriminate objects from background, is more suitable for the recognition task than that of the STL method, which is generative in nature and aims to capture how objects look in general. More detailed comparison on specific object recognition tasks is presented in Figure 3.

On computational cost, STL is very expensive compared to KAT for high-dimensional data. This is mainly due to the costly  $l_1$ -norm sparsity optimization in STL. On a computer server with an Intel dual-core 2.93GHz CPU and 24GB RAM, it took between 4 to 30 hours to learn STL sparse bases depending on the sparsity weight and about 5-20 minutes per category for computing the coefficients of all training and testing data. In contrast, our KAT took on average 20 minutes to learn each category so is at least one magnitude faster.

**Recognition using only auxiliary knowledge.** To evaluate the significance of selecting unlabelled auxiliary knowledge in KAT based on the limited target class samples, we compare KAT to a generic non-background classifier termed ‘Aux’ using auxiliary data only. Table 1 shows that based on the general knowledge

Target Category	Manually Selected Auxiliary Categories	Non-transfer	KAT	Supervised KAT	Target Category	Manually Selected Auxiliary Categories	Non-transfer	KAT	Supervised KAT
	    	0.353	<b>0.341</b>	0.388		    	0.443	0.430	<b>0.317</b>
	    	0.47	<b>0.38</b>	0.43		    	0.406	0.425	<b>0.329</b>
	    	0.398	<b>0.372</b>	0.416		    	0.356	<b>0.327</b>	0.396

Fig. 5. Supervised KAT vs. KAT.

#Target Samples	#Pos / #Neg				Avg EER					
	KAT	KAT ( $\lambda = 0$ )	KAT-Naive	KAT (full sparsity)	KAT ( $\lambda = 0$ )	KAT-Naive	KAT (full sparsity)	Kernel Alignment	Non-Transfer	
1	203 / 84	203 / 84	196 / 92	196 / 92	0.392	0.392	0.401	0.401	0.398	0.428
2	208 / 65	203 / 78	166 / 114	151 / 128	0.370	0.371	0.399	0.416	0.414	0.404
5	187 / 84	134 / 138	108 / 184	84 / 205	0.357	0.387	0.424	0.446	0.442	0.373
10	66 / 25	72 / 80	22 / 138	25 / 137	0.343	0.354	0.413	0.415	0.382	0.349
15	112 / 28	76 / 151	44 / 196	36 / 204	0.325	0.363	0.428	0.450	0.403	0.336

Table 2. Further investigation on transferrable knowledge selection.

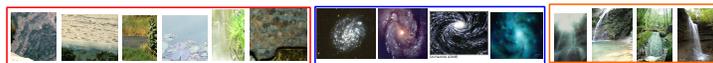
extracted from all objects in the auxiliary dataset alone, the performance of Aux is much better than random guess (0.5), but worse than the non-transfer model and much worse than our KAT. This result confirms the necessity and usefulness of performing transfer learning provided that the extracted auxiliary knowledge is indeed useful, and is quantified and selected.

**Supervised KAT vs. KAT.** To compare how unsupervised transfer learning fares against supervised transfer learning given a fully labelled auxiliary dataset of only related object categories to a target class, we selected related auxiliary categories for 6 target categories (car tire, giraffe, lathe, cormorant, gorilla, and fire truck) as shown in Figure 5. With these labelled and related auxiliary data, we implemented a supervised version of our KAT, that is, auxiliary mappings were constructed from each auxiliary class, instead of relying on clustering. As shown in Figure 5, with those labelled data from the related auxiliary categories, the supervised KAT does not always achieve better results compared to (unsupervised) KAT. This is because our KAT can select automatically those related samples for building auxiliary kernels when they are present, but more important also utilises any relevant and available information about how objects are distinguishable from background from all the irrelevant auxiliary data.

#### Further investigation on transferrable knowledge selection.

1) *With vs. without constraints.* We validate the usefulness of the two constraints introduced in KAT: (1) a hypothesis constraint (Eqn. (8)) and (2) a partial sparsity penalty controlled by  $\lambda$  (Eqn. (9)), for alleviating negative transfer. Table 2 shows the performance of KAT without the second constraint (termed KAT( $\lambda = 0$ )) and without both (termed KAT-Naive). Note that as aforementioned for one target sample, the  $\lambda$  in KAT as well as its variants was set to 0. It is evident that the usefulness of the extracted auxiliary knowledge is significantly weakened (more than 33% higher in average EER compared to KAT in some case) without these constraints and much less positive transfer is obtained.

2) *KAT vs. Kernel Alignment.* We compare KAT with the widely adopted kernel alignment method for kernel fusion in Table 2 [23]. As shown KAT performs much better, because KAT introduces our two proposed constraints to



**Fig. 6.** Examples of object categories that no improvement was obtained using KAT. From left to right: background, galaxy, and waterfall.

balance the effects of target and auxiliary kernels, but kernel alignment does not explicitly consider the important difference between those two effects.

3) *KAT vs. KAT(full sparsity)*. We show that the most popular way for kernel selection using a full  $l_1$  norm sparsity on all kernel weights  $b_i, i = 0, \dots, m$  cannot work well for our unsupervised cross-category transfer learning. This is demonstrated by the KAT(full sparsity) that uses the full sparsity as in previous MKL work [29] on all kernel weights for kernel fusion in Table 2. The results show that without differentiating target kernel from auxiliary kernels using the proposed constraints, the full sparsity penalty leads to much worse performance.

**The failure mode.** As shown in table 1, some object categories cannot benefit from KAT. Figure 6 shows examples of three such categories. For these categories, the sample images either contain large portion of background (e.g. waterfall) or contain objects that do not have clear contour but have similar textures as cluttered background (e.g. galaxy). Since the similarity to background is greater than that to other object classes in auxiliary data (see Figure 4), the extracted transferrable knowledge thus would not help for these object categories.

## 6 Conclusion

We introduced a novel unsupervised selective transfer learning method using Kernel Adaptation Transfer (KAT) which utilises unlabelled auxiliary data of largely irrelevant object classes to any target object category. The model quantifies and selects the most relevant transferrable knowledge for recognising any given target object class with very limited samples. For one-class recognition, KAT selects the most useful general knowledge about how objects are visually distinguishable from cluttered background. Our experiments demonstrate clearly that due to its transferrable knowledge selection capability, the proposed unsupervised KAT model significantly outperforms the Self-Taught Learning (STL) method. Despite only implemented for one-class recognition tasks in this work, the proposed KAT framework is a general transfer learning method that can be readily formulated for other pattern recognition problems with large intra-class variation and sparse training data conditioned that a set of auxiliary mappings  $f_s$  can be constructed. Our current work includes extending this model to multi-class object detection and one-to-many object verification.

**Acknowledgement.** This research was partially funded by the EU FP7 project SAMURAI with grant no. 217899.

## References

1. Viola, P., Jones, M.: Rapid object detection using a boosted cascade of simple features. In: CVPR. (2001)
2. Pinker, S.: How the minds works (1999) W. W. Norton.

3. Bart, E., Ullman, S.: Cross-generalization: Learning novel classes from a single example by feature replacement. In: CVPR. (2005)
4. Farhadi, A., Endres, I., Hoiem, D., Forsyth, D.: Describing objects by their attributes. In: CVPR. (2009)
5. Lampert, C.H., Nickisch, H., Harmeling, S.: Learning to detect unseen object classes by between-class attribute transfer. In: CVPR. (2009)
6. Stark, M., Goesele, M., Schiele, B.: A shape-based object class model for knowledge transfer. In: ICCV. (2009)
7. Fei-Fei, L., Fergus, R., Perona, P.: One-shot learning of object categories. PAMI **28** (2006) 594–611
8. Pan, S.J., Yang, Q.: A survey on transfer learning. IEEE Transactions on Knowledge and Data Engineering **22** (2010) 1345–1359
9. Zweig, A., Weinshall, D.: Exploiting object hierarchy: Combining models from different category levels. In: ICCV. (2007)
10. Pan, S.J., Kwok, J.T., Yang, Q.: Transfer learning via dimensionality reduction. In: AAAI. (2008) 677C682
11. Pan, S.J., Tsang, I.W., Kwok, J.T., Yang, Q.: Domain adaptation via transfer component analysis. In: IJCAI. (2009)
12. Yang, J., Yan, R., Hauptmann, A.G.: Cross-domain video concept detection using adaptive svms. In: ACM Multimedia. (2008) 188–197
13. Duan, L., Tsang, I.W., Xu, D., Maybank, S.J.: Domain transfer svm for video concept detection. In: CVPR. (2009)
14. Rosenstein, M.T., Marx, Z., Kaelbling, L.P., Dietterich, T.G.: To transfer or not to transfer. In: NIPS 2005 Workshop on Transfer Learning. (2005)
15. Griffin, G., Holub, A., Perona, P.: Caltech-256 object category dataset (2007) Technical Report UCB/CSD-04-1366, California Institute of Technology.
16. Fergus, R., Perona, P., Zisserman, A.: Object class recognition by unsupervised scale-invariant learnings. In: CVPR. (2003)
17. Raina, R., Battle, A., Lee, H., Packer, B., Ng, A.Y.: Self-taught learning: transfer learning from unlabeled data. In: ICML. (2007) 759–766
18. Dai, W., Jin, O., Xue, G.R., Yang, Q., Yu, Y.: Eigentransfer: a unified framework for transfer learning. In: ICML. (2009) 25
19. Lanckriet, G., Cristianini, N., Bartlett, P., Ghaoui, L., Jordan, M.: Learning the kernel matrix with semidefinite programming. JMLR **5** (2004) 27–72
20. Gehler, P., Nowozin, S.: On feature combination for multiclass object classification. In: ICCV. (2009)
21. Daumé, H.: Frustratingly easy domain adaptation. In: ACL. (2007)
22. Nocedal, J., Wright, S.: Numerical optimization (2006) 2nd ed., Springer.
23. Cristianini, N., Kandola, J., Elisseeff, A., Shawe-Taylor, J.: On kernel-target alignment. In: NIPS. (2002)
24. Hillel, A.B., Hertz, T., Weinshall, D.: Efficient learning of relational object class models. In: ICCV. (2005)
25. Kadir, T., Brady, M.: Saliency and image description. IJCV **45** (2001) 83–105
26. Lowe, D.: Distinctive image features from scale-invariant keypoints. IJCV **2** (2004) 91–110
27. Crandall, D.J., Huttenlocher, D.P.: Weakly supervised learning of part-based spatial models for visual object recognition. In: ECCV. (2006)
28. Schnitzspan, P., Fritz, M., Schiele, B.: Hierarchical support vector random fields: Joint training to combine local and global features. In: ECCV. (2008)
29. Rakotomamonjy, A., Bach, F., Canu, S., Grandvalet, Y.: More efficiency in multiple kernel learning. In: ICML. (2007)