The more you know, the less you learn: from knowledge transfer to one-shot learning of object categories

Tatiana Tommasi
http://www.idiap.ch/~ttommasi/
Idiap Research Institute
Martigny, CH

Barbara Caputo
http://people.idiap.ch/caputo/
Ecole Polytechnique Federale EPFL
Lausanne, CH

Abstract

Learning a category from few examples is a challenging task for vision algorithms, while psychological studies have shown that humans are able to generalise correctly even from a single instance (one-shot learning). The most accredited hypothesis is that humans are able to exploit prior knowledge when learning a new related category. This paper presents an SVM-based model adaptation algorithm able to perform knowledge transfer for a new category when very limited examples are available. Using a leave-one-out estimate of the weighted error-rate the algorithm automatically decides from where to transfer (on which known category to rely), how much to transfer (the degree of adaptation) and if it is worth transferring something at all. Moreover a weighted least-squares loss function takes optimally care of data unbalance between negative and positive examples. Experiments presented on two different object category databases show that the proposed method is able to exploit previous knowledge avoiding negative transfer. The overall classification performance is increased compared to what would be achieved by starting from scratch. Furthermore as the number of already learned categories grows, the algorithm is able to learn a new category from one sample with increasing precision, i.e. it is able to perform one-shot learning.

1 Introduction

A major goal in object categorisation is learning and recognising effectively thousands of categories, as humans do [1]. To this end, a very promising trend is to develop methods for learning from small samples by exploiting prior experience via knowledge transfer. The basic intuition is that, if a system has already learned N categories, learning the $N+1^{th}$ should be easier, even from one or few training samples, because the algorithm can take advantage of what was learned already [21]. When considering knowledge transfer approaches to object categorisation, it is worth keeping in mind the following issues: *(a) when to transfer:* while intuitively one might assume that prior knowledge is going to help in learning a new category, this might not always be the case. Consider for instance a system that has learned so far only different categories of animals (dogs, cats, ducks, dolphins etc). When it starts to learn the new category "motorbike", it is not obvious that the prior knowledge is going to help much. The ideal knowledge transfer algorithm should be able to determine

automatically if it is worthwhile transferring knowledge or not; *(b) from where to transfer:* we would expect that knowledge transfer will be more effective between similar categories. For instance, when learning from few samples the category motorbike, it would help more to transfer knowledge from models of other types of vehicles (cars, trucks, etc) rather than from models of animals. This means having an algorithm able to measure quantitatively the similarity between a new category and all the old ones stored in memory, and to use this information for determining from where to transfer.

Several approaches have been proposed so far for transferring knowledge, spanning from transferring model parameters [6, 7, 12, 19], to samples [10, 11, 14, 23], to general categorical properties [8], using also information coming from unlabelled data [17, 18]. While all of these approaches proved to work reasonably well in some domain, how to transfer is still an open research question. We argue that an ideal algorithm should transfer knowledge so to boost learning when only one/few samples are available (the so called "one-shot learning" phenomenon). The one-shot learning effect should become stronger as the number of known categories grows, because in that case it is most likely that the system has already learned a category very similar to the one to be learned.

This paper presents an algorithm that addresses these issues. We take a discriminative approach, and we cast the object categorisation problem in a Least Square-Support Vector Machine (LS-SVM, [4]) framework. We build on recent work on LS-SVM-based model adaptation [16], where a crucial requirement is having available many samples of the new class. Here we extend the model in order to enable it to learn a new category even from only one image. The resulting algorithm determines automatically from where to transfer and how much to rely on the transferred knowledge. Also, thorough experiments on two different databases show that, when the number of known categories grows, the performance obtained by using only one training image increases dramatically, clearly showing a one-shot learning effect.

In the rest of the paper we review LS-SVM, describe the model adaptation method presented in [16] and derive our knowledge transfer approach (Section 2). Experiments showing the power of our algorithm are presented in Section 3. We conclude with an overall discussion and plans for future work.

2 The Knowledge Transfer Learning Approach

Let us suppose to have a category detection algorithm that has been trained so far to recognise N categories. This concretely corresponds to define N functions $f_j(\mathbf{x}) \rightarrow \{1,-1\}, j = 1,...,N$ such that the image $\mathbf{x}$ is assigned to the j^{th} category if and only if $f_j(\mathbf{x}) = 1$. When beginning to learn the $N+1^{th}$ category, the algorithm will have initially only one/few samples for learning $f_{N+1}(\mathbf{x})$. Our goal is to exploit, whenever possible, the existing prior knowledge to boost the learning of $f_{N+1}(\mathbf{x})$. In the following we will briefly review the LS-SVM theory (Section 2.1) and how it can be used in a model adaptation framework [16] (Section 2.2). Starting from this, we will show how it is possible to derive a knowledge transfer algorithm able to determine automatically when and where from to transfer, with a one-shot learning behaviour in presence of a rich prior knowledge (Section 2.3).

2.1 Least Square-Support Vector Machine

Let us assume to have a binary problem and a set of l samples $\{\mathbf{x}_i, y_i\}_{i=1}^{l}$ where $\mathbf{x}_i \in \mathscr{X} \subset \mathbb{R}^d$ is an input vector describing the i^{th} sample and $y_i \in \mathscr{Y} = \{-1,1\}$ is its label. The

goal of the SVM classifier is to learn a linear model that assigns the correct label to an unseen test sample [4]. This can be thought as learning a linear function $f(\mathbf{x}) = \mathbf{w} \cdot \phi(\mathbf{x}) + b$ where $\phi(\mathbf{x})$ maps the input samples to a high dimensional feature space, induced by a kernel function $K(\mathbf{x}, \mathbf{x}') = \phi(\mathbf{x}) \cdot \phi(\mathbf{x}')$. In LS-SVM the model parameters $(\mathbf{w}, b)$ are found solving the following constrained optimisation problem [4]:

$$\min_{\mathbf{w},b} \frac{1}{2}\|\mathbf{w}\|^2 + \frac{C}{2}\sum_{i=1}^{l}\xi_i^2 \quad \text{subject to} \quad y_i = \mathbf{w}\cdot\phi(\mathbf{x}_i) - b + \xi_i \quad \forall\, i \in \{1,\cdots,l\}. \tag{1}$$

The corresponding primal Lagrangian is [4]:

$$\mathscr{L} = \frac{1}{2}\|\mathbf{w}\|^2 + \frac{C}{2}\sum_{i=1}^{l}\xi_i^2 - \sum_{i=1}^{l}\alpha_i\{\mathbf{w}\cdot\phi(\mathbf{x}_i) + b + \xi_i - y_i\}, \tag{2}$$

where $\boldsymbol{\alpha} = (\alpha_1, \alpha_2, \ldots, \alpha_l) \in \mathbb{R}^l$ is a vector of Lagrange multipliers. The optimality conditions for the obtained problem define a system of linear equations that can be written concisely in matrix form as [3]:

$$\begin{bmatrix} \mathbf{K} + \frac{1}{C}\mathbf{I} & \mathbf{1} \\ \mathbf{1}^T & 0 \end{bmatrix} \begin{bmatrix} \boldsymbol{\alpha} \\ b \end{bmatrix} = \begin{bmatrix} \mathbf{y} \\ 0 \end{bmatrix} \tag{3}$$

where $\mathbf{K}$ is the kernel matrix. Let us call $\mathbf{G}$ the first left-hand side matrix in (3). It turns out that the least-square optimisation problem can be solved by simply inverting $\mathbf{G}$.

The accuracy of the model on test data is critically dependent on the choice of good learning parameters (e.g. the kernel parameters and the regularization parameter C). This choice can be based on a preliminary cross validation evaluating the leave-one-out error, which is known to be approximately an unbiased estimator of the classifier generalisation error [15]. LS-SVM allows to write the leave-one-out error $r_i^{(-i)}$ for the i^{th} sample in closed form [3]. Let $[\alpha^{(-i)}; b^{(-i)}]$ represent the dual parameters of the LS-SVM when the i^{th} sample is omitted during the leave-one-out cross validation procedure. It is shown that [3]: $[\alpha^{(-i)}; b^{(-i)}] = \mathbf{G}_{(-i)}^{-1}[y_1, \ldots, y_{i-1}, y_{i+1}, \ldots, y_l, 0]^T$, where $\mathbf{G}_{(-i)}$ is the matrix obtained when the i^{th} sample is omitted in $\mathbf{G}$. Using the block matrix invertion lemma we have [3]:

$$r_i^{(-i)} = \frac{\alpha_i}{\mathbf{G}_{ii}^{-1}}. \tag{4}$$

So without explicitly running cross validation experiments it is possible to define a criterion error to maximise the LS-SVM model generalisation performance [3]:

$$ERR = \sum_{i=1}^{l}\Psi\{y_i r_i^{(-i)} - 1\} \quad \text{with} \quad \Psi\{z\} = \frac{1}{1 + exp\{-10 * z\}}, \tag{5}$$

the best learning parameters are those minimising this error.

2.2 Learning a new object category from many samples

Let us assume that we want to learn a new category from a set of labelled training data $\{\mathbf{x}_i\}_{i=1,m}$, taking advantage of what learned so far. Orabona et al. [16] proposes to start the training with a known model and then refine it through adaptation. Adaptation is defined

constraining a new model to be close to one of a set of pre-trained models. The proposed method is mathematically formulated in the LS-SVM classification framework changing the classical regularization term and defining the following optimisation problem [16]:

$$\min_{\mathbf{w},b} \frac{1}{2}\|\mathbf{w}-\beta\mathbf{w}'\|^2+\frac{C}{2}\sum_{i=1}^{l}\xi_i^2 \quad \text{subject to} \quad y_i=\mathbf{w}\cdot\phi(\mathbf{x}_i)-b+\xi_i \quad \forall i\in\{1,\cdots,l\} \tag{6}$$

where $\mathbf{w}'$ is the parameter describing the old model and β is a scaling factor necessary to control the degree to which the new model is close to the old one. The optimal solution [16]:

$$\mathbf{w}=\beta\mathbf{w}'+\sum_{i=1}^{l}\alpha_i\phi(\mathbf{x}_i), \tag{7}$$

is given by the sum of the pre-trained model scaled by the parameter β and a new model built on the new data points. When β is 0 the obtained formula comes back to the original LS-SVM formulation, that is without any adaptation to the previous data. To find the optimal β, the authors take advantage from the possibility of LS-SVM to write the leave-one-out error in closed form. It turns out that it is still possible to do it for the modified formulation in (6) obtaining:

$$r_i^{(-i)}=\frac{\alpha_i}{\mathbf{G}_{ii}^{-1}}-\beta\frac{\alpha_i'}{\mathbf{G}_{ii}^{-1}} \tag{8}$$

where $\alpha_i'=\mathbf{G}_{(-i)}^{-1}[\hat{y}_1,\ldots,\hat{y}_{i-1},\hat{y}_{i+1},\ldots,\hat{y}_l,0]^T$ and $\hat{y}_i=(\mathbf{w}'\cdot\phi(\mathbf{x}_i))$, i.e. $\hat{y}_i$ is the prediction of the old model on the i^{th} sample. The obtained leave-one-out error depends on β, so for each known model it is possible to find the best β producing the lowest criterion error ERR (5). Moreover, comparing all the criterion errors, the lowest one identifies the best prior knowledge model to use for adaptation.

We call this algorithm *Adapt*, it was proposed for learning adaptively grasping postures for prosthetic hands [16] and seems very promising also for learning new object categories with knowledge transfer. The model from where to transfer is chosen as the one producing the lowest criterion error, and knowledge is transferred in the form of its model parameter $\mathbf{w}'$. The scaling factor β determines how much to transfer, again depending on the criterion error evaluation. Note that all of this is learned automatically by the algorithm. A major drawback is that when learning from less than 150 samples, results are unstable, due to the high variance of the leave-one-out error technique when considering few samples. In the next section we will show that we overcome this point by introducing weighting factors that "rebalance" the problem and that makes it possible to use effectively this method even when learning from one single image.

2.3 Learning a new object category from few samples

Suppose to have a training set with 1 positive and 20 negative examples, on the basis of which we want to estimate from where to transfer, using the leave-one-out error. Making a wrong prediction on one of the examples contributes for 1/20 of the total error independently respect to the sign of its label. This is not good: we would like to be more tolerant on negative examples due to their higher number, and strict on the positive one which is alone. In such cases, to use effectively the criterion error, it is necessary to reweight the leave-one-out

recognition of positive and negative examples. A way to do it is to modify the criterion error to have a leave-one-out cross-validation estimate of the Weighted Error Rate (WERR) [3]:

$$WERR = \sum_{i=1}^{l} \zeta_i \Psi\{y_i r_i^{(-i)} - 1\} \quad \text{where} \quad \zeta_i = \begin{cases} \frac{l}{2l^+} & \text{if } y_i = +1 \\ \frac{l}{2l^-} & \text{if } y_i = -1. \end{cases} \tag{9}$$

Here the function Ψ is the same as in (5) and l^+ and l^- represent the number of positive and negative examples respectively. Introducing the weighting factors ζ_i is asymptotically equivalent to re-sampling the data so that object and non-object samples are balanced [3]. If we consider again a training set with 1 positive and 20 negative examples, the introduction of the described weight makes the error on a negative example contribute for 1/40 of the total while the error on the positive example contribute for 1/2. Let us identify with *Adapt-W* the adaptation method described in the previous Section with ERR (5) substituted by WERR (9).

As already mentioned, the WERR helps in the selection of the best prior knowledge and in defining its relevance for the new task. This means that it gives a contribution just on the "final" part of the knowledge transfer method, but not while building the new adapted model. To take care of the data unbalance also during this "initial" step, we propose to find the model parameters $(\mathbf{w}, b)$ via minimisation of a regularised weighted least-square loss function [20]:

$$L = \frac{1}{2}\|\mathbf{w}\|^2 + \frac{C}{2}\sum_{i=1}^{l} \zeta_i [y_i - \mathbf{w}\cdot\phi(\mathbf{x}_i) - b]^2. \tag{10}$$

It introduces just a small variation in the LS-SVM solution: the optimal dual model parameters $(\boldsymbol{\alpha}, b)$ are defined by a modified system of linear equations [3]:

$$\begin{bmatrix} \mathbf{K} + \frac{1}{C}\mathbf{W} & \mathbf{1} \\ \mathbf{1}^T & 0 \end{bmatrix} \begin{bmatrix} \boldsymbol{\alpha} \\ b \end{bmatrix} = \begin{bmatrix} \mathbf{y} \\ 0 \end{bmatrix} \tag{11}$$

where $\mathbf{W} = diag\{\zeta_1^{-1}, \zeta_2^{-1}, \ldots, \zeta_l^{-1}\}$ and ζ_i are defined as in (9). Let's call the obtained variant *LS-SVM-W*.

Hence the model adaptation method can be changed to its weighted formulation:

$$\min_{\mathbf{w},b} \frac{1}{2}\|\mathbf{w} - \beta\mathbf{w}'\|^2 + \frac{C}{2}\sum_{i=1}^{l} \zeta_i \xi_i^2 \quad \text{subject to} \quad y_i = \mathbf{w}\cdot\phi(\mathbf{x}_i) - b + \xi_i \quad \forall i \in \{1, \cdots, l\}. \tag{12}$$

In this way the weighting factors ζ_i take into account that the proportion of positive and negative examples in the training data are known not to be representative of the operational class frequencies. More in detail, the ξ_i term represents the misclassification cost of the i-th datum during training. Here, introducing a weight let the classification model to be built balancing the contribution of penalties coming from different labelled examples. In the case of 1 positive and 20 negative examples, the misfit ξ_i term is multiplied by a factor 1/40 for a negative sample and 1/2 for the positive one. Let's call *Adapt-2W* the strategy which combines together the weighted model adaptation technique (12) and the WERR (9). In this way we define a new knowledge transfer method which allows to learn new visual categories from few examples as shown by our experimental results.

3 Experiments

We present here three set of experiments, designed for studying the behaviour of our algorithm when (a) it knows few categories, and none of them is very similar to the new one

(unrelated categories, Section 3.2); (b) it knows few categories that are very similar to the new one (related categories, Section 3.3), and (c) the number of known categories increases, with a specific focus on the one-shot performance (Section 3.4). The experiments were run on two subsets of two different object category databases: the Caltech-256 [9] and the IRMA database used in the CLEF challenge 2008 [5]. In the rest of this Section we first describe the experimental setup (Section 3.1), and then we report our findings for the three scenarios described above.

3.1 Experimental setup

Our working assumption is to have N category detection models stored in memory, built using standard SVM and looking for the optimal $\mathbf{w}'$. We used the Gaussian kernel $K(\mathbf{x},\mathbf{x}') = exp(-\gamma\|\mathbf{x}-\mathbf{x}'\|^2)$ [4] for all experiments; the parameters C and γ were chosen by cross-validation. When the new $N+1$ object category comes, the system starts learning. The new data consists of m images from a background dataset and an increasing number of instances of the new category, from 1 to m. Each experiment is repeated on five different ordering of the data, chosen randomly. Moreover, to get a reliable estimate of the adaptation performance on all the considered categories, we used a leave-one-out approach, using in turn each class for adaptive learning and considering all the remaining categories as prior knowledge. At each step the performance is evaluated on an equal number of unseen background images and instances of the new category. The parameters C and γ for the adaptive LS-SVM were chosen as described above for the known categories, and only the scaling factor β was selected through the leave-one-out cross validation estimate of WERR (9).

In the following we will compare the performance of *Adapt-W* to that of *Adapt-2W*. Moreover we consider the performance of *LS-SVM* and *LS-SVM-W* trained only on the new incoming data, which correspond respectively to (6) and (12) where $\beta = 0$. We do not directly compare against *Adapt* [16], because it does not work on small training samples. We now describe the experimental setup specific to the two chosen databases.

Caltech 256 setup We considered eight object categories from the Caltech-256 database [9], namely bulldozer, car-side, firetruck, motorbike, schoolbus, snowmobile, dog and duck. From the original dataset, for each category, we selected images where the object was clearly visible and where it always had the same orientation. This resulted in datasets with a minimum of 33 images (schoolbus) and a maximum of 83 images (snowmobile). We used the whole category clutter (827 images), randomly selecting a background class for each category. As features we used the Pyramid Histograms of Oriented Gradients (PHOG) [2]. We computed descriptors with orientation in the range [0,360] and we built a histogram with K=8 bins. We considered $L = 3$ levels in forming the pyramid grid [2]. The resulting feature vector has 680 elements.

IRMA setup The IRMA database[1] is a collection of radiographs presenting a large number of rich classes defined according to a four-axis hierarchical code [13]. We decided to work on the 2008 IRMA database version [5], just considering the third axis of the code: it describes the anatomy, namely which part of the body is depicted, independently to the used acquisition technique or direction. 23 classes with more than 100 images were selected from various sub-levels of the third axis, 3 of them were used to define the background class[2]. As

[1] Available from http:\\phobos.imib.rwth-aachen.de\irma\datasets_en.php.

[2] 213-nose area (242 images), 230-neuro area (365 images), 310-cervical spine (508 images), 320-thoracic spine (279 images), 330-lumbar spine (540 images), 411-hand finger (325 images), 414- left hand (541 images), 415-right hand (176 images), 421-left carpal joint (124 images), 441-left elbow (114 images), 442-right elbow (105 images),

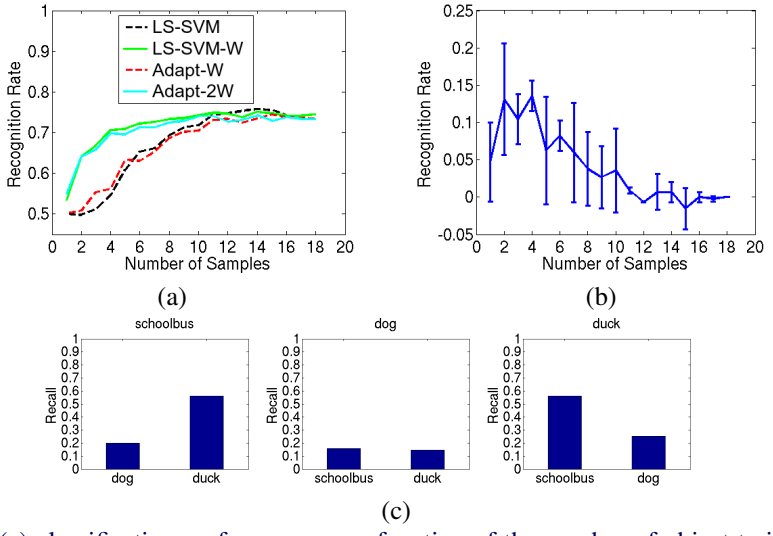


Figure 1: (a) classification performance as a function of the number of object training images, when learning three unrelated categories. The results showed correspond to average recognition rate over the three categories, considering each class-out experiment repeated 5 times. (b) average difference in classification performance $\pm$ standard deviation, obtained by *Adapt-2W* with respect to *Adapt-W*. (c) for each class-out experiment, the histogram bars represent the known categories recall on the test set, indicating the prior knowledge capability in recognising the new object.

features we used the SIFT-based approach described in [22].

3.2 Experiments on unrelated categories

In the first set of experiments we considered three visually different categories to understand if the adaptation model is negatively affected by transferring from unrelated tasks. We chose schoolbus, dog and duck from the described dataset and from each category we selected randomly 36 images for training (18 object and 18 background instances) and 30 images for testing (15 object and 15 background instances). Results are showed in Figure 1(a): we see that the *Adapt-W* and *LS-SVM* curves are almost identical as well as *Adapt-2W* and *LS-SVM-W*: if the WERR evaluation does not indicate any of the known classes as helpful, both adaptation methods perform roughly as the corresponding non adaptative methods. Moreover we see that *Adapt-2W* performs better than *Adapt-W*: Figure 1(b) shows that *Adapt-2W* has an improvement of up to 14% in recognition rate for less than 10 object images compared to *Adapt-W*. The two methods asymptotically coincide. Figure 1(c) shows, for each category, the average recall of the known classes on the test set. These results can give an intuition about the reliability of the known categories for the new task. It is clear that in each case there is very few useful information stored in memory.

463-right humero-scapular joint (146 images), 610-right breast (144 images), 620-left breast (155 images), 914-left foot(146 images), 915-right foot (139 images), 921-left ankle joint (192 images), 922-right ankle joint (229 images), 942-left knee (231 images), 943-right knee (222 images). Three classes used for background: 700-abdomen (219 images), 800-pelvis (263 images), 500-chest (4611 images).

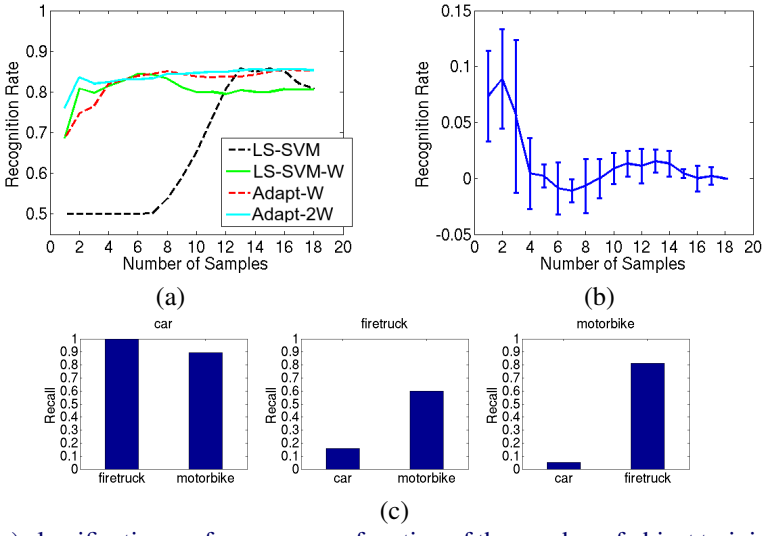
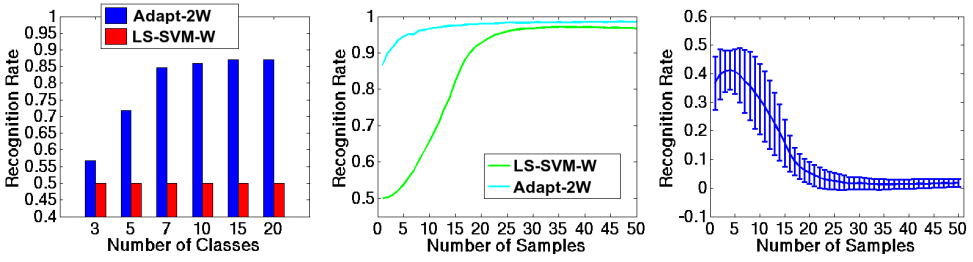


Figure 2: (a) classification performance as a function of the number of object training images when learning three related categories. The results showed correspond to average recognition rate over the three categories, considering each class-out experiment repeated 5 times. (b) average difference in classification performance $\pm$ standard deviation obtained by the *Adapt-2W* method with respect to *Adapt-W*. (c) for each class-out experiment, the histogram bars represent the known categories recall on the test set, indicating the prior knowledge capability in recognising the new object.

3.3 Experiments on related categories

In the second set of experiments we considered three visually related categories, all belonging to the Caltech-256 general class "motorized transportation" [9]. We chose car, firetruck and motorbike from the described dataset and from each we selected randomly 36 images for training (18 object and 18 background instances) and 30 images for testing (15 object and 15 background instances). From Figure 2(a) we can see that adaptation produces clearly better results than starting from scratch. Moreover the difference in recognition rate showed in Figure 2(b) indicate that by using *Adapt-2W* we have an improvement in recognition rate of up to 9% for less than 4 object images in the training set, compared to using *Adapt-W*. Finally, Figure 2(c) shows for each category the average recall of the prior knowledge classes. This indicate that in each case there is at least one good known reliable category to use for adaptation. The same set of experiments was repeated considering all the six visually related categories in our dataset (bulldozer, car, firetruck, motorbike, schoolbus and snowmobile) from the Caltech-256 general class "motorized transportation" [9]. The obtained results are similar to what showed on three categories: using *Adapt-2W* we have better results (up to 5% in recognition rate) for less than 5 object images in the training set, compared to using *Adapt-W*, while the two methods asymptotically coincide. Moreover it is possible to notice that the one-shot learning performance is improved respect to the three class case. For *Adapt-2W* the recognition rate using only one object instance in the training set goes from 76% for three categories to 79% for six categories.



(a) (b) (c)

Figure 3: (a) one-shot learning performance of the *Adapt-2W* and corresponding *LS-SVM-W*, varying the total number of categories. (b) classification performance as a function of the number of training images when learning on 20 object categories. The results correspond to average recognition rate over the 20 categories, considering each class-out experiment repeated 5 times. (c) average classification performance difference obtained by the *Adapt-2W* method with respect to *LS-SVM-W*. The error bars denote $\pm$ standard deviation with respect to the average values.

3.4 Experiments on an increasing number of categories

All the experimental results showed till here asses the higher performance of the *Adapt-2W* respect to the *Adapt-W* method. For this reason we decided to use just the first approach for the experiments on the IRMA database. Here we study how performance varies when the number of known categories grows. We are especially interested in monitoring how the method behaves when learning from one single image. We randomly selected from each category 100 instances for training and 100 instances for testing (for both the sets, 50 object and 50 background images). Five sets of experiments were run considering 3/5/7/10 and 15 classes plus a final one with all the 20 categories. We started extracting three categories through random selection and then we went on adding new ones till covering the whole 20 class dataset. Figure 3(a) shows the obtained recognition rate results for *Adapt-2W* and the corresponding *LS-SVM-W* when only one object image is used for training. We expect that the overall performance of the knowledge transfer method will increase along with the number of stored models, since there is a larger probability to find a matching pre-trained model. This intuition is confirmed by the increasing trend in the one-shot learning recognition rate. This trend is quite fast at the beginning passing from 3 (57% recognition rate) to 5 (72% recognition rate) and 7 (85% recognition rate) categories and then becomes slower from 10 (86% recognition rate) to 20 categories (both for 15 and 20 classes the one-shot learning rate is 87%). We show in Figure 3(b) the 20 categories results and in 3(c) the corresponding difference in performance when using the adaptation method with respect to learning form scratch. As one can see, adaptation uniformly obtains a better performance showing an asymptotic gain of about 2.5%.

4 Conclusions and Future work

We presented an SVM-based method for learning object categories from few examples using knowledge transfer. The algorithm decides automatically from where and how much to transfer, adapting the known model to the incoming data. The reliability of prior knowledge for the new task is evaluated by estimating its generalisation error so to weight properly pos-

itive and negative examples in the training set. Moreover the model adaptation is appropriately designed to balance the possible misfit of object and non-object instances. Experiments show that the proposed method improves the learning performance when useful information is stored in memory, while it never affects it negatively when the known categories are very different from the new one. When the number of known categories increases, the performance of the model improves remarkably, showing a one-shot learning behaviour. In the future we plan to run experiments to understand more deeply the algorithm capabilities and to compare with the results presented in [8]. Moreover, we would like to extend the method to multiple cues, and to hierarchical categorisation, with the aim to reduce the computational complexity of the algorithm for large number of known categories.

Acknowledgments

This work was supported by the EU project DIRAC (FP6-0027787) and by the EMMA project thanks to the Hasler foundation (www.haslerstiftung.ch). We are thankful to Francesco Orabona and to the anonymous reviewers for their many helpful comments and suggestions.

References

[1] I. Biederman. Recognition-by-components: A theory of human image understanding. *Psychological Review*, 94.

[2] A. Bosch, A. Zisserman, and X. Munoz. Representing shape with a spatial pyramid kernel. In *CIVR '07: Proceedings of the 6th ACM international conference on Image and video retrieval*, 2007.

[3] G.C. Cawley. Leave-one-out cross-validation based model selection criteria for weighted ls-svms. In *proceedings IJCNN*, Vancouver, Canada, July 2006.

[4] N. Cristianini and J. Shawe-Taylor. *An Introduction to Support Vector Machines (and Other Kernel-Based Learning Methods)*. Cambridge University Press, 2000.

[5] T. Deselaers and T. Deserno. Medical image annotation in imageclef 2008. In *working notes CLEF*, 2008.

[6] E.Bonilla, K.M. Chai, and C. Williams. Multi-task gaussian process prediction. In *Proceedings of NIPS*, 2008.

[7] T. Evgeniou and M. Pontil. Regularized multi-task learning. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2004.

[8] Li Fei-Fei, R. Fergus, and P. Perona. One-shot learning of object categories. *IEEE Transactions on Pattern Analysis and Machine Intelligenc*, 28.

[9] G. Griffin, A. Holub, and P. Perona. Caltech 256 object category dataset. Technical Report UCB/CSD-04-1366, California Institue of Technology, 2007.

[10] J. Huang, A. Smola, A. Gretton, K.M. Borgwardt, and B. Schölkopf. Correcting sample selection bias by unlabeled data. In *Proceedings of NIPS*, 2007.

[11] J.Jiang and C. Zhai. Instance weighting for domain adaptation in nlp. In *Proceedings of the Annual Meeting of the Association of Computational Linguistics*, 2007.

[12] N.D. Lawrence and J.C. Platt. Learning to learn with the informative vector machine. In *Proceedings of ICML*, 2004.

[13] T.M. Lehmann, H.Schubert, D. Keysers, M. Kohnen, and B.B. Wein. The irma code for unique classification of medical images. In *Proceedings SPIE*, 2003.

[14] X. Liao, Y. Xue, and L. Carin. Logistic regression with an auxiliary data source. In *Proceedings of ICML*, 2005.

[15] A. Lunz and V. Brailovsky. *On estimation of characters obtained in statistical procedure of recognition (in russian)*, volume 3. Techicheskaya Kibernetica, 1969.

[16] F. Orabona, C. Castellini, B. Caputo, E. Fiorilla, and G. Sandini. Model adaptation with least-squares svm for hand prosthetics. In *proceedings ICRA*, 2009.

[17] A. Quattoni, M. Collins, and T. Darrell. Transfer learning for image classification with sparse prototype representations. *Computer Vision and Pattern Recognition, IEEE Computer Society Conference on*, 0:1–8, 2008.

[18] R. Raina, A. Battle, H. Lee, B. Packer, and A. Y. Ng. Self-taught learning: transfer learning from unlabeled data. In *ICML '07: Proceedings of the 24th international conference on Machine learning*, pages 759–766. ACM, 2007.

[19] A. Schwaighofer, V. Tresp, and K. Yu. Learning gaussian process kernels via hierarchical bayes. In *Proceedings of NIPS*, 2005.

[20] J.A.K. Suykens, T. Van Gestel, J. De Brabanter, B. De Moor, and J. Vanderwalle. *Least Squares Support Vector Machines*. World Scientific, 2002.

[21] S. Thrun. Is learning the n-th thing any easier than learning the first? In *Advances in Neural Information Processing Systems*, pages 640–646. The MIT Press, 1996.

[22] T. Tommasi, F. Orabona, and B. Caputo. An svm confidence-based approach to medical image annotation. In *Evaluating Systems for Multilingual and Multimodal Information Access – Proceedings of CLEF*, 2008.

[23] B. Zadrozny. Learning and evaluating classifiers under sample selection bias. In *Proceedings of ICML*, 2004.