Detection and Identification of Rare Audiovisual Cues

Insperata accident magis saepe quam quae speres.
(Things you do not expect happen more often than things you do expect) Plautus (ca 200(B.C.))

Project no: 027787

DIRAC

Detection and Identification of Rare Audio-visual Cues

Integrated Project
IST - Priority 2

DELIVERABLE NO: D4.14
TRANSFER OF KNOWLEDGE FOR LEARNING OF RARE EVENTS

Date of deliverable: 30.6.2010
Actual submission date: 12.8.2010

Start date of project: 01.01.2006 Duration: 60 months

Organization name of lead contractor for this deliverable: **IDIAP Research Institute**

Revision [0]

Project co-funded by the European Commission within the Sixth Framework Program (2002-2006)		
Dissemination Level		
PU	Public	
PP	Restricted to other program participants (including the Commission Services)	
RE	Restricted to a group specified by the consortium (including the Commission Services)	X
CO	Confidential, only for members of the consortium (including the Commission Services)	

Detection and Identification of Rare Audiovisual Cues

Insperata accident magis saepe quam quae speres.
(Things you do not expect happen more often than
things you do expect) Plautus (ca 200(B.C.))

D4.14 TRANSFER OF KNOWLEDGE FOR LEARNING OF RARE EVENTS

IDIAP Research Institute

Abstract:

Within the context of detection of incongruent events, an often overlooked aspect is how a system should react to the detection. The set of all the possible actions is certainly conditioned by the task at hand, and by the embodiment of the artificial cognitive system under consideration. Still, a desirable action that does not depend from these factors is to update the internal model and learn the new detected event. This calls for algorithms able to learn a new class from few labeled samples, as the data available for learning a rare event is, by definition, very little.

In this document we present an algorithm for knowledge transfer that determines automatically on which known category to rely (from where to transfer), the degree of adaptation (how much to transfer) and if it is worth transferring something at all. A preliminary version of the algorithm has been presented at the British Machine Vision Conference in 2009. A generalized version of the method has been presented at the International Conference on Computer Vision and Pattern Recognition in 2010.

Table of Content

1. Introduction

The capability to recognize, and react to, rare events is one of the key features of biological cognitive systems. In spite of its importance, the topic is little researched. The DIRAC framework defines rareness as an incongruence compared to the prior knowledge of the system. Concretely, this means organizing the prior knowledge in layers of general and specific classifiers. Whenever a specific classifier rejects, but its corresponding general classifier accepts, the system is in presence of an incongruent event.

A still almost completely unexplored aspect of the framework is how to react to the detection of an incongruent event. Of course, this is largely influenced by the task at hand, and by the type of embodiment of the artificial system under consideration: the type of reactions that a camera might have are bound to be different from the type of actions a wheeled robot might take. Still, there is one action that is desirable for every system, regardless of their given task and embodiment: to learn the detected incongruent event, so to be able to recognize it correctly if encountered again in the future.

Here we propose a new transfer learning method as a suitable candidate for learning y newly detected incongruent event. The method is able to learn a new class from few, even one single labeled example by exploiting optimally the prior knowledge of the system. This corresponds, in the DIRAC framework, to transfer from the general class that has accepted.

We focus on inductive transfer learning. In order to produce a model with generalization capabilities, a learning algorithm must have an inductive bias – a set of assumptions about the true distribution of the training data. Our method is based on Least Square SVM and consists into learning a new task via adaptation. Prior knowledge is defined by the learned models, and is transferred constraining a new model to be close to a weighted set of pre-trained models.

A first version of the method, allowing transfer learning for a single pre-trained model, was presented at the British Machine Vision Conference in 2009. An extension of the model, allowing for transfer learning from multiple pre-trained models, was presented at the International Conference on Computer Vision and Pattern Recognition in 2010. In the rest of this document we provide the re-prints of these two publications.

2. Reference

[1] T. Tommasi, B. Caputo. The more you know, the less you learn: from knowledge transfer to one-shot learning of object categories. *Proceedings of the British Machine Vision Conference –BMVC 2009,* 2009.

[2] T. Tommasi, F. Orabona, B. Caputo.Safety in numbers: learning categories from few examples with multi model knowledge transfer. *Proceedings of the International Conference on Computer Vision and Pattern Recognition –CVPR 2010,* 2010.

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, 17, 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 [10]:

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

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 [7]. 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 [7], 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),

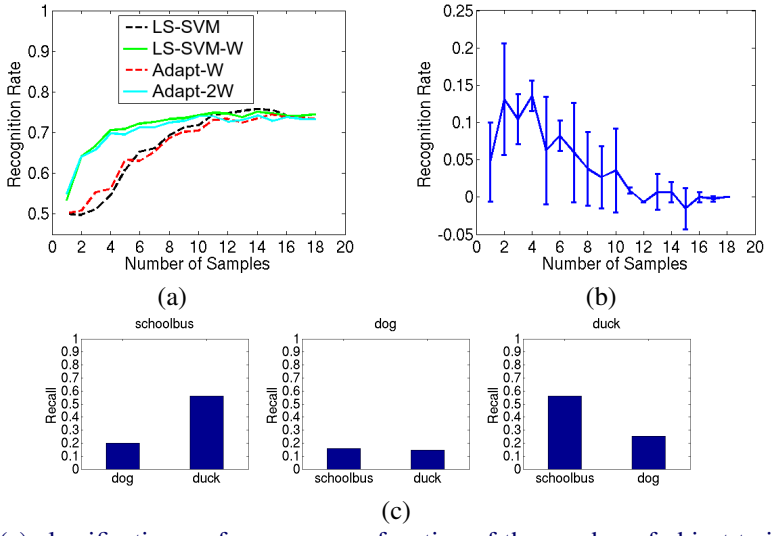


Figure 1: (a) classification performance as a function of the number of object training images, when learning three unrelated categories. The results showed correspond to average recognition rate over the three categories, considering each class-out experiment repeated 5 times. (b) average difference 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).

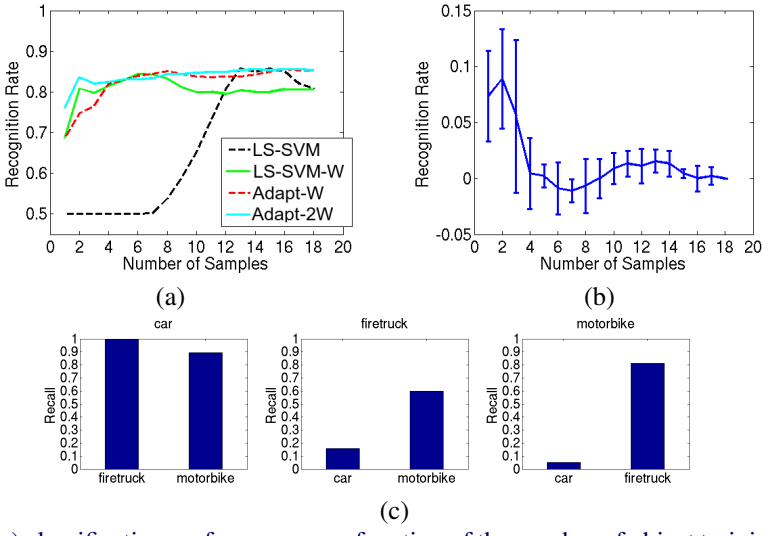
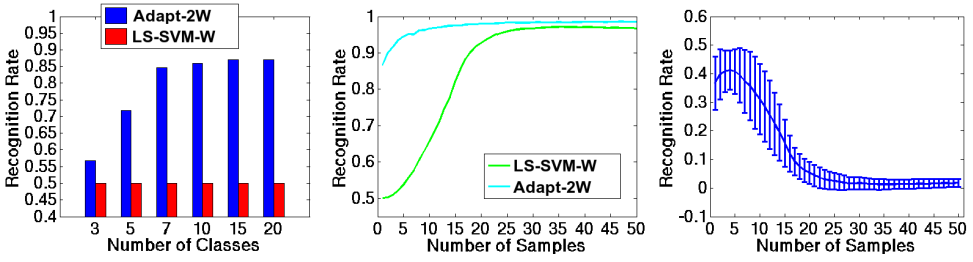


Figure 2: (a) classification performance as a function of the number of object training images when learning three related categories. The results showed correspond to average recognition rate over the three categories, considering each class-out experiment repeated 5 times. (b) average difference in 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.

Safety in Numbers: Learning Categories from Few Examples with Multi Model Knowledge Transfer

Tatiana Tommasi[1,2] , Francesco Orabona[3] , Barbara Caputo[1]
[1]Idiap Research Institute, Martigny CH
[2]Ecole Polytechnique Federale de Lausanne, EPFL, Lausanne CH
[3]University of Milan, Milan IT
ttommasi@idiap.ch, ..., bcaputo@idiap.ch

Abstract

Learning object categories from small samples is a challenging problem, where machine learning tools can in general provide very few guarantees. Exploiting prior knowledge may be useful to reproduce the human capability of recognizing objects even from only one single view. This paper presents an SVM-based model adaptation algorithm able to select and weight appropriately prior knowledge coming from different categories. The method relies on the solution of a convex optimization problem which ensures to have the minimal leave-one-out error on the training set. Experiments on a subset of the Caltech-256 database show that the proposed method produces better results than both choosing one single prior model, and transferring from all previous experience in a flat uninformative way.

1. Introduction

The ability to learn from few samples is a hallmark of human intelligence. We rapidly and reliably learn many kinds of regularities and this enables us to make inductive inference even from only small amount of data [1].

Although current state of the art categorization methods reach impressive results on difficult datasets [6], they don't handle well small training sets. Without additional information, learning from few examples always reduces to an ill-posed optimization problem. A possible solution is exploiting prior knowledge, a strategy known in the literature as *learning to learn, knowledge transfer* or *transfer learning*. The basic intuition is that, if a system has already learned k categories, learning the $k+1$ should be easier, even from one or few training samples [21]. Besides boosting the learning process, knowledge transfer can give three other advantages ([19], see Figure 1): *(1) higher start:* the initial performance is higher (one-shot learning); *(2) higher slope:* performance grows faster, and *(3) higher asymptote:* the final performance is better.

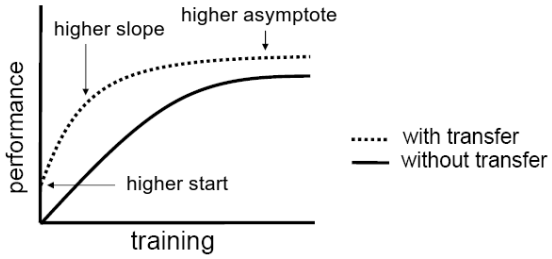


Figure 1. Three ways in which transfer might improve learning [19].

The contribution of this paper is a method for learning object categories from few examples. We focus on three key issues for knowledge transfer: *how to transfer, what to transfer* and *when to transfer* [16]. We propose a discriminative method based on Least Square Support Vector Machine (LS-SVM)[20] (*how to transfer*) that learns the new class through adaptation. We define the prior knowledge as the hyperplanes of the classifiers $\mathbf{w}'_j$, $j = 1, \cdots, k$ of the k classes already learned (*what to transfer*). Hence knowledge transfer is equivalent to constrain the hyperplanes $\mathbf{w}$ of the $k+1$ new category to be close to those of a sub-set of the k classes. This strategy is in between the choice of transferring acritically from all previously learned models [7] and transferring from one single model [22]. We learn the sub-set of classes from where to transfer, and how much to transfer from each of them, via the Leave One Out (LOO) error on the training set. Determining how much to transfer helps avoiding negative transfer. Therefore, in case of non-informative prior knowledge, transfer might be disregarded completely (*when to transfer*).

Experiments on various subsets of the Caltech-256 [12] database show that our approach consistently reproduces the curve depicted in Figure 1 with a higher start, and higher slope compared to what is obtained by not exploiting prior knowledge, and to current state of the art knowledge tran-

sfer approaches [22, 7]. Furthermore, when the number k of known classes grows, our algorithm presents a one-shot learning behaviour.

The rest of the paper is organized as follows: we give an overview of previous work in Section 2. Section 3 describes LS-SVM and the knowledge transfer algorithm in [22], on which we build. Section 4 describes our new algorithm and discusses its properties. Experimental results are reported in Section 5. We conclude with an overall discussion and pointing out possible avenues for future research.

2. Related Work

Several authors addressed in the past the issues of what, how and when to transfer. We review below the most prominent approaches.

What to Transfer. We can find three answers to this question in the literature (see [16] for a survey). The first is the *instance-transfer approach*: although the source domain data cannot be reused directly, there are certain parts of them that can still be considered together with a few labelled data in the target domain. A second solution is defined by *transferring feature representations*. It means learning a common feature structure, *e.g.* a kernel in SVM-based approaches, from different domains that can bridge related tasks. The third strategy can be described as *parameter-transfer approach*. It assumes that the source task and the target tasks share some parameters of their model or priors.

How to Transfer. Wu and Dietterich transferred source training examples either as support vectors or as constraints (or both) and demonstrated improved image classification by SVMs [24]. Fei-Fei *et al*. proposed a Bayesian transfer learning approach for object recognition [7] that learns a common prior over visual classifier parameters. Zweig *et al*. [25] investigated transfer learning with a method based on combining object classifiers from different hierarchical levels into a single classifier. Using discriminative (maximum margin) object models, Fink developed a method that learns distance metrics from related problems [9]. Quattoni *et al*. [17] proposed to use knowledge transfer in an unsupervised setting learning a representation based on kernel distances to available unlabelled data.

When to Transfer. Works focusing on when to transfer evaluate the limit of transfer learning power. Rosenstain *et al*. [18] showed empirically that if two tasks are dissimilar, then the transferring hurts the performance on the target task. Ideally, a transfer method should produce positive transfer between appropriately related tasks while avoiding negative transfer when the tasks are not a good match. However it might be easier to avoid negative transfer if, given multiple source tasks, one transfers from several or all of them.

Research on knowledge transfer is still in its infancy, especially applied to object recognition. Although there are many publication in this area, given the slightly different tasks defined in each paper, none of them compare against the others. Moreover there is not an official testbed database, nor a standard experimental setup. In this work we propose a reproducible experimental setting that can be used in the future to test new knowledge transfer algorithms and we benchmark our algorithm against two other methods in literature [22, 7]. We address three open problems of [8]: (1) The possibility that a sophisticated multimodal prior is beneficial in learning; (2) if it is easier to learn new categories which are similar to some of the "prior" categories; (3) if exist another point of view besides the Bayesian one that allows to incorporate prior knowledge. We present a discriminative method which exploits a combination of multiple visual features and selects automatically the most useful prior knowledge models to use when learning a new category.

3. Problem Statement

Consider the following scenario. We have k visual categories and a classifier trained to distinguish each of them from background. This corresponds to define k functions $f_j(\mathbf{x}) \rightarrow \{1, -1\}, j = 1, ..., k$ such that the image $\mathbf{x}$ is assigned to the j^{th} category if and only if $f_j(\mathbf{x}) = 1$. Now suppose that we want to learn a new $k+1$ category from just one or few instances, plus some background examples. To obtain f_{k+1} we can train using only the available data, or we can take advantage of what already learned. In the following we briefly review the LS-SVM theory and how it can be used in a model adaptation framework [15]. We review how this approach can be formulated to derive a knowledge transfer algorithm that exploits prior knowledge from *only one* of the k classes [22] (Section 3.1). The contribution of this paper is how to extend this method to exploit *all* the suitable prior knowledge. The used strategy is presented in Section 4.

3.1. LS-SVM Adaptation Method: learning from small samples

Suppose to have a binary problem and a set of l samples $\{\mathbf{x}_i, y_i\}_{i=1}^{l}$ where $\mathbf{x}_i \in \mathcal{X} \subset \mathbb{R}^d$ is an input vector describing the i^{th} sample and $y_i \in \mathcal{Y} = \{-1, 1\}$ is its label. We want to learn a linear function $f(\mathbf{x}) = \mathbf{w} \cdot \phi(\mathbf{x}) + b$ which assigns the correct label to an unseen test sample $\mathbf{x}$. $\phi(\mathbf{x})$ is used to map 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 by solving the following optimisation problem:

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

It can be shown [20] that the optimal $\mathbf{w}$ is expressed by $\mathbf{w} = \sum_{i=1}^{l} \alpha_i \phi(\mathbf{x}_i)$, and $(\boldsymbol{\alpha}, b)$ are found solving

$$\begin{bmatrix} \mathbf{K} + \frac{1}{C}\mathbf{I} & \mathbf{1} \\ \mathbf{1}^T & 0 \end{bmatrix} \begin{bmatrix} \boldsymbol{\alpha} \\ b \end{bmatrix} = \begin{bmatrix} \mathbf{y} \\ 0 \end{bmatrix}, \quad (2)$$

where $\mathbf{K}$ is the kernel matrix. Let us call $\mathbf{G}$ the first term in left-hand side of (2). The least-square optimisation problem can be solved by simply inverting $\mathbf{G}$. Another advantage of the LS-SVM formulation is that it gives the possibility to write the LOO error in closed form [4]. The LOO error is an unbiased estimator of the classifier generalization error and can be used for model selection.

Slightly changing the classical LS-SVM regularization term, it is possible to define a learning method based on adaptation [15]. The idea is to constrain a new model to be close to one of a set of pre-trained models:

$$\min_{\mathbf{w},b} \frac{1}{2}\|\mathbf{w} - \beta\mathbf{w}'\|^2 + \frac{C}{2}\sum_{i=1}^{l}[y_i - \mathbf{w}\cdot\phi(\mathbf{x}_i) - b]^2, \quad (3)$$

where $\mathbf{w}'$ is the parameter describing the old model and β is a scaling factor in $(0, 1)$ necessary to control the degree to which the new model is close to the old one. The LOO error in the modified formulation is:

$$r_i^{(-i)} = \frac{\alpha_i}{\mathbf{G}_{ii}^{-1}} - \beta\frac{\alpha_i'}{\mathbf{G}_{ii}^{-1}}, \quad (4)$$

where $\alpha_i' = \mathbf{G}_{(-i)}^{-1}[\hat{y}_1, \ldots, \hat{y}_{i-1}, \hat{y}_{i+1}, \ldots, \hat{y}_l, 0]^T$, $\mathbf{G}_{(-i)}$ is the matrix obtained when the i^{th} sample is omitted in $\mathbf{G}$ 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. $r_i^{(-i)}$ is then used to obtain an estimate of the Weighted Error Rate (WER) [4]:

$$WER = \sum_{i=1}^{l} \zeta_i \Psi\{y_i r_i^{(-i)} - 1\} \quad (5)$$

$$\text{with} \quad \Psi\{z\} = \frac{1}{1 + exp\{-10 * z\}} \quad (6)$$

$$\text{and} \quad \zeta_i = \begin{cases} \frac{l}{2l^+} & \text{if } y_i = +1 \\ \frac{l}{2l^-} & \text{if } y_i = -1\,. \end{cases} \quad (7)$$

Here 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 examples are balanced [4]. Hence, without explicitly running cross validation experiments, the best learning parameters which maximise the LS-SVM model generalisation performance can be found as those minimising WER. Since $r_i^{(-i)}$ depends on β, for each known model it is possible to find the best β producing the lowest WER. Then, comparing all the criterion errors, the lowest one will identify the best prior knowledge model to use for adaptation.

To further increase robustness to unbalanced distributions of the data, the model parameters $(\mathbf{w}, b)$ can be found via minimisation of a regularised weighted least-square loss function [20]:

$$\min_{\mathbf{w},b} \frac{1}{2}\|\mathbf{w}\|^2 + \frac{C}{2}\sum_{i=1}^{l}\zeta_i[y_i - \mathbf{w}\cdot\phi(\mathbf{x}_i) - b]^2. \quad (8)$$

This introduces just a small variation in the LS-SVM solution: the optimal model parameters $(\boldsymbol{\alpha}, b)$ are defined by a modified system of linear equations [4]:

$$\begin{bmatrix} \mathbf{K} + \frac{1}{C}\mathbf{W} & \mathbf{1} \\ \mathbf{1}^T & 0 \end{bmatrix} \begin{bmatrix} \boldsymbol{\alpha} \\ b \end{bmatrix} = \begin{bmatrix} \mathbf{y} \\ 0 \end{bmatrix}, \quad (9)$$

where $\mathbf{W} = diag\{\zeta_1^{-1}, \zeta_2^{-1}, \ldots, \zeta_l^{-1}\}$ and ζ_i are defined as in (7). Hence the model adaptation method changes to its weighted formulation [22]:

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

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. In particular they help to balance the contribution of the sets of positive and negative examples to the data misfit term [22].

Experiments show that this method is able to learn new visual categories from few examples. However, the algorithm can choose only one prior known model. As we will show in Section 5, this is not always the best solution. Moreover this approach can suffer for instability in time, *i.e.* when the number of training images increases.

4. Multi Model Knowledge Transfer

Consider the following situation. Suppose to be given the task to learn from few examples the class motorbike, having already learned the categories bicycle, car, dog and cat. We would expect to achieve better results by transferring from bicycle *and* car, rather than transferring from bicycle *or* car. Also, we would expect better results compared to transferring equally from *all* known categories, as cat and dog might induce negative transfer.

This kind of scenario motivates us to design a knowledge transfer algorithm able to find autonomously the best subset of known models from where to transfer. In the rest of the Section we define the new model (Section 4.1) and we discuss its properties (Section 4.2).

4.1. Multi Model Knowledge Transfer: Definition

We start from Equation (10) and we rewrite it substituting the single coefficient β with a vector $\boldsymbol{\beta}$ containing as many elements as the number of prior models, k:

$$\min_{\mathbf{w},b} \frac{1}{2} \left\| \mathbf{w} - \sum_{j=1}^{k} \beta_j \mathbf{w}'_j \right\|^2 + \frac{C}{2} \sum_{i=1}^{l} \zeta_i (y_i - \mathbf{w} \cdot \phi(\mathbf{x}_i) - b)^2 \,. \tag{11}$$

Here $\boldsymbol{\beta}$ has to be chosen in the unitary ball, *i.e.* $\|\boldsymbol{\beta}\|_2 \leq 1$. It is similar to the regularization term used in LS-SVM in Equation (1), and it is a natural generalization of the original constraint $0 \leq \beta \leq 1$. This term is necessary to avoid overfitting problems. They can happen when the number of known models is large compared to the number of training samples. With this new formulation the optimal solution is

$$\mathbf{w} = \sum_{j=1}^{k} \beta_j \mathbf{w'}_j + \sum_{i=1}^{l} \alpha_i \phi(\mathbf{x}_i) \,. \tag{12}$$

Hence $\mathbf{w}$ is expressed as a weighted sum of the pre-trained models scaled by the parameters β_j, plus the new model built on the incoming training data.

To find the optimal $\boldsymbol{\beta}$ we use again the possibility of LS-SVM to write the LOO error in closed form:

$$r_i^{(-i)} = y_i - \tilde{y}_i = \frac{\alpha_i}{\mathbf{G}^{-1}_{ii}} - \sum_{j=1}^{k} \beta_j \frac{\alpha'_{i(j)}}{\mathbf{G}^{-1}_{ii}}, \tag{13}$$

where $\alpha'_{i(j)} = \mathbf{G}^{-1}_{(-i)}[\hat{y}^j_1, \ldots, \hat{y}^j_{i-1}, \hat{y}^j_{i+1}, \ldots, \hat{y}^j_l, 0]^T$, $\hat{y}^j_i = (\mathbf{w}'_\mathbf{j} \cdot \phi(\mathbf{x}_i))$ and $\tilde{y}$ is the LOO prediction. By multiplying everything by y_i we obtain:

$$y_i \tilde{y}_i = 1 - y_i \left(\frac{\alpha_i}{\mathbf{G}^{-1}_{ii}} - \sum_{j=1}^{k} \beta_j \frac{\alpha'_{i(j)}}{\mathbf{G}^{-1}_{ii}} \right) \,. \tag{14}$$

The best values of β_j are those minimizing the LOO error, *i.e.* the values producing positive values for $y_i \tilde{y}_i$, for each i. However minimizing directly the sign of those quantities would result in a non-convex formulation with many local minima. We propose instead the following loss function:

$$\begin{aligned} loss(y_i, \tilde{y}_i) &= \zeta_i \max\left[1 - y_i \tilde{y}_i, 0\right] \\ &= \max\left[y_i \zeta_i \left(\frac{\alpha_i}{\mathbf{G}^{-1}_{ii}} - \sum_{j=1}^{k} \beta_j \frac{\alpha'_{i(j)}}{\mathbf{G}^{-1}_{ii}} \right), 0 \right] \,. \end{aligned} \tag{15}$$

This loss function is similar to the hinge loss used in Support Vector Machines. It is a convex upper bound to the LOO misclassification loss and favours solution in which $\tilde{y}_i$ has a value of 1, beside having the same sign of y_i. Moreover it has a smoothing effect, similar to the function in (6).

Finally, the objective function is:

$$J = \sum_{i=1}^{l} loss(y_i, \tilde{y}_i) \quad \text{s.t.} \quad \|\boldsymbol{\beta}\|_2 \leq 1 \,. \tag{16}$$

Notice that this formulation is equivalent to the more common optimization problem $(1/2)\|\boldsymbol{\beta}\|_2^2 + CJ$ for a proper choice of C [5]. By minimizing J we can find the best values of β_j to weight the known prior models in the transfer learning process. The scaling factors ζ_i are introduced in the loss function to take care of the data unbalance between positive and negative samples in the training set, as in [22].

We implement the optimization process using a simple projected sub-gradient descent algorithm, where at each iteration $\boldsymbol{\beta}$ is projected onto the l_2-sphere, $\|\boldsymbol{\beta}\|_2 \leq 1$.

4.2. Multi Model Knowledge Transfer: Properties

The main advantage of our approach is the ability to transfer from multiple prior model, instead of choosing just one. At the same time, the knowledge is not transfered in a flat, uninformative way, but we evaluate the importance of each model and their interaction. The loss used is convex and the constraint in (16) is convex too, hence the minimizer of (16) is unique. This is opposed to the formulation proposed in [22], where (7) is non-convex. This means that the algorithm in [22] can have many local minima.

An important property of this new formulation is also its "stability". Stability here means that the behaviour of the algorithm does not change much if a point is removed or added. This notion is closely related to the LOO error, which is exactly calculated measuring the performance of the model every time a point is removed. Recent works have shown that a stable algorithm has a better generalization ability [3]. The algorithm in [22] can choose only one model at each time step, to be used to transfer knowledge. This means that everytime the algorithm "changes its mind", *i.e.* it chooses a different prior model on two consecutive time steps, the behaviour of the algorithm will change completely. On the other hand, our method selects more than one prior model at each time step, so we expect that differences between steps in the vector $\boldsymbol{\beta}$ will be small. The regularization is also important in this sense [3]. In Section 5.2 we show empirically that this is true.

From a computational point of view the current algorithm's runtime is $O(l^3 + kl^2)$, with l the number of training samples (of the order of 10 images) and k the number of known prior models. The first term is related to inverting $\mathbf{G}$, while the second term is the computational complexity of (13). We match the complexity of a plain SVM, which in the worst case is known to be $O(l^3)$ [13], and is the standard out-of-the-shelf classification method commonly used on datasets with more than 10^3 images. The computational complexity of each step of the projected sub-gradient

descent is $O(kl)$ and it is extremely fast. For instance, our MATLAB implementation takes just half a second with $l = 12$ and $k = 3$.

5. Experiments

We present here three sets of experiments designed to illustrate how our algorithm performs (a) when the prior knowledge is related/unrelated to the new class (Section 5.2) (b) when prior knowledge increases (Section 5.3) (c) compared to the current state of the art [7, 8] (Section 5.4) We first describe the experimental setup (Section 5.1) and then we report our findings in the three scenarios described above.

5.1. Experimental Setting

Our working assumption is to have k category models stored in memory, built using LS-SVM. We used the Gaussian kernel $K(\mathbf{x}, \mathbf{x}') = \exp(-\gamma\|\mathbf{x} - \mathbf{x}'\|^2)$ for all our experiments; the parameters C and γ were chosen by cross-validation. When the new $k+1$ category comes, the system starts learning.

All experiments are run on subsets of the Caltech-256 database [12]. We selected in total 41 classes + background class, obtaining a data set with a fair amount of clutter and scale variation. We didn't perform any image selection or preprocessing. The training data consists of m images from the background dataset and an increasing number of instances of the new category from 1 to m. The test set consists of 100 images, half from the background and half from the new category. Images are chosen randomly by splitting each class into two disjoint sets: m training images are drawn randomly for the first, a set of 50 are taken from the second. As we focus on learning from small samples, we varied m from 1 to 6, repeating the experiments 10 times for each value and using different sets of training and test images. To get a reliable estimate of the performance on all the categories, we used a leave-one-class-out approach, considering in turn each class for adaptive learning and using all the rest as prior knowledge.

We used the pre-computed features of [10] which the authors made available on their website[1]. Specifically, we used four different image descriptors: PHOG Shape Descriptors [2], Appearance Descriptors [14], Region Covariance [23] and Local Binary Patterns. All of the image descriptors were computed in a spatial pyramid, we considered just the first level (*i.e.* informations extracted from the whole image) and combined the features using the average kernel.

In the following we will compare the performance of our Multi Model Knowledge Transfer aglorithm (*Multi-KT*) to that obtained with a flat average mixture of prior knowledge (*Average-KT*) and to the method presented in [22] that we call here *Single-KT*.We also benchmark all the results against *No Adapt*. This corresponds to learn from scratch using weighted-LS-SVM, *i.e.* solving the optimization problem in Equation (10) with $\beta = 0$. The significance of the comparisons are evaluated through the sign test [11].

[1] http://www.vision.ee.ethz.ch/~pgehler/projects/iccv09/

5.2. Related/Unrelated Prior Knowledge

In the first set of experiments we considered different groups of related and unrelated categories. The goal is to study how *Multi-KT* chooses the reliable prior knowledge, and its impact on performance.

Related Classes. We considered two sets of 6 classes belonging respectively to the Caltech-256 general classes "transportation, ground, motorized" (bulldozer, firetruck, motorbikes, schoolbus, snowmobile, car-side) and "food edibles" (cake, hamburger, hot-dog, ice-cream-cone, spaghetti, sushi). Figure 2(a)-(d) show the respective classification results. In both cases we see that all KT algorithms obtain an impressive advantage over starting from scratch. As Figure 2(c)-(f) shows, *Multi-KT* performs clearly better than *Single-KT*, with ($p < 0.02$) for less than four images in both cases. This confirms the intuition that it pays off to transfer from multiple sources, as opposed to one, when they all bring useful information. There is no significant difference in accuracy between *Multi-KT* and *Average-KT* (Figure 2(b)-(e)). This suggests that, when all prior knowledge is useful, learning the weights does not give a real advantage over a flat average.

Mixed Classes. To consider what happens in a more confused situation, we selected the following 10 classes: dog, horse, zebra, helicopter, fighter-jet, motorbikes, car-side, dolphin, goose and cactus. The classification results in Figure 2(g) show that here *Multi-KT* performs better both than *Average-KT* and than *Single-KT* (Figure 2(h)-(i), in both cases $p < 0.02$ for less than four images). This experiment illustrates very clearly the power of our approach: when the prior knowledge is partially related to the new class, transferring from only one model does not exploit fully previous experience. At the same time, using acritically all the prior knowledge induces partial negative transfer behaviours, that affect the overall performance. Notice that the situation of knowledge transfer from a mixture of related and unrelated classes is the most common.

We can also compare *Multi-KT* to *Single-KT* in terms of stability. Let us consider the unique β used by *Single-KT* as an element of the $\boldsymbol{\beta}$ vector where all the remaining elements are zero. There are 6 steps in time corresponding to a new positive sample entering the training set. For each couple of subsequent steps we calculated the difference between the obtained $\boldsymbol{\beta}$ vectors of *Single-KT*. We did the same with the $\boldsymbol{\beta}$ vectors produced by the *Multi-KT* algorithm. Figure 3

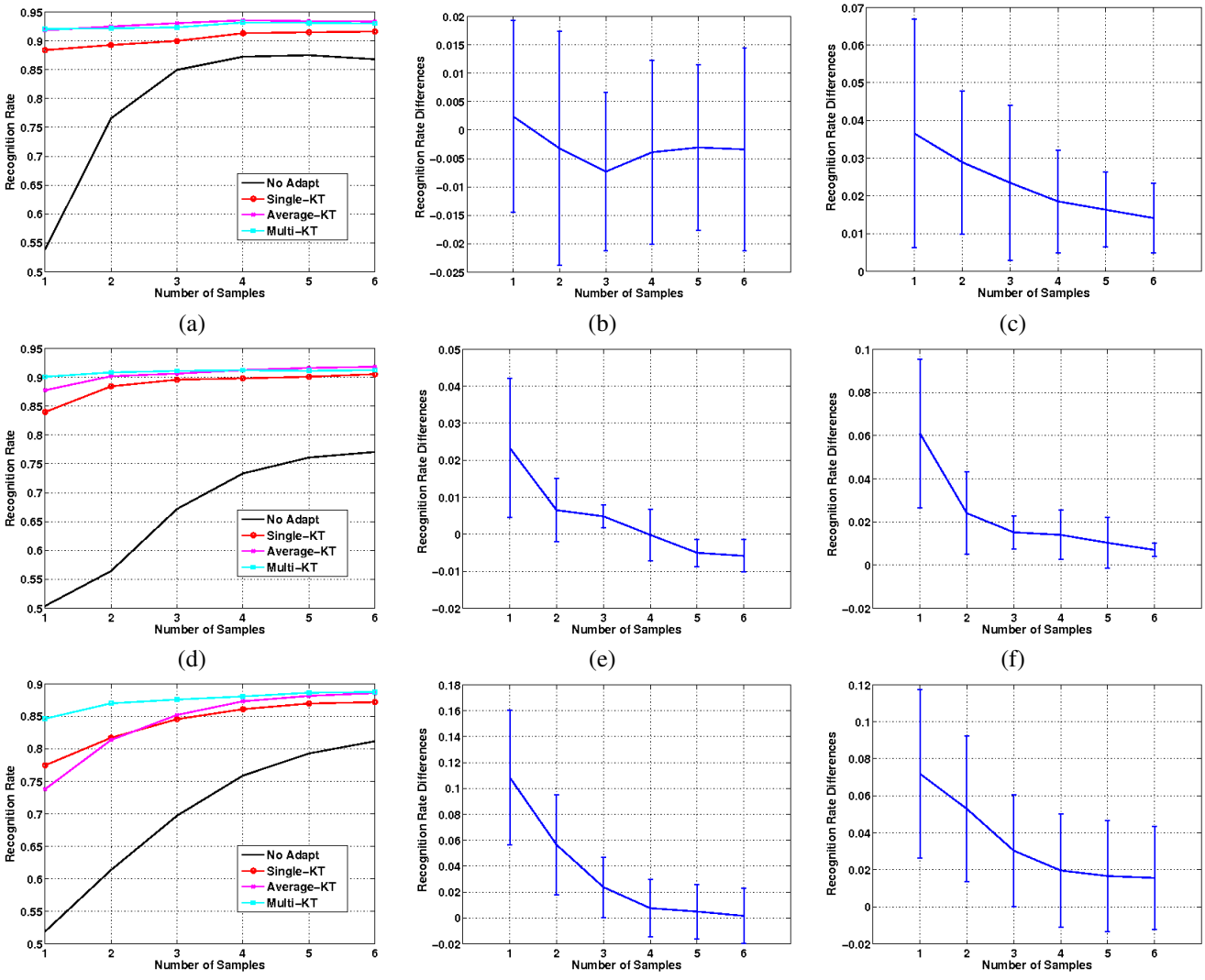


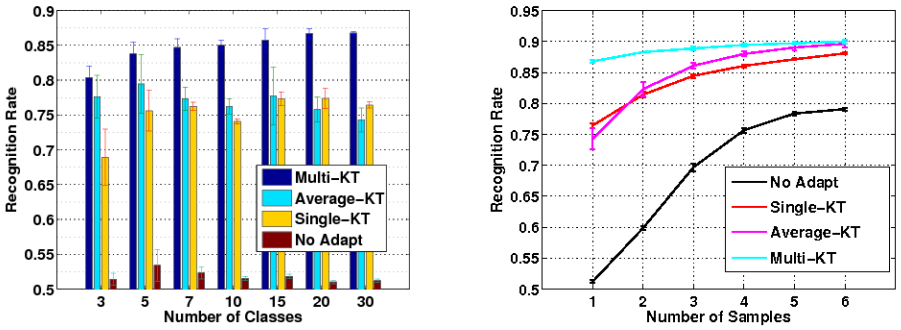
Figure 2. (a-d-g) Classification performance as a function of the number of object training images, when learning respectively one out of six related categories "transportation, ground, motorized", "food, edibles", and one out of ten mixed categories. The results shown correspond to average recognition rate over the categories, considering each class-out experiment repeated 10 times; (b-e-h) average difference in classification performance $\pm$ standard deviation, obtained by *Multi-KT* with respect to *Average-KT*; (c-f-i) average difference in classification performance $\pm$ standard deviation, obtained by *Multi-KT* with respect to *Single-KT*.

shows the average norm of these differences. It is evident that choosing a combination of the prior known models for transfer learning is more stable (lower average variations in the $\boldsymbol{\beta}$ vectors) than relying on just a single known category.

5.3. Increasing Prior Knowledge

Here we studied how performance varies when the number of known category grows. We are especially interested in how *Multi-KT* behaves when learning from a single positive image (one-shot learning). We selected 30[2] classes, extracting 3 visually related classes from 10 general categories of Caltech-256. We run six set of experiments, considering 3/5/7/10/15/20 categories plus a final one with all the 30 categories. We first extracted three categories through random selection and then we went on adding new ones till covering the whole 30 class dataset. We repeated the experiments three times: Figure 4(a) shows the average recognition rate and the corresponding standard deviations when training only on one object image. We expect that the overall performance will increase along with the num-

[2]"transportation, ground, motorized":car-side, fire-truck, motorbike; "animal,land": dog, horse, zebra; "animal,water": goldfish, dolphin, killer-whale; "transportation, water": canoe, kayak, speed-boat; "music, stringed": electric-guitar, harp, mandolin; "food, containers": beer-mug, coffee-mug, teapot; "transportation, air": airplanes, helicopter, fighter-jet; "animals, air"': duck, goose, swan; "plants": bonsai, cactus, fern; "structures, buildings": light-house, windmill, smokestack.



(a) (b)

Figure 4. (a) One-shot learning performance of *Multi-KT*, *Average-KT* and *Single-KT* respect to *No Adapt* when varying the number of prior known categories; (b) classification performance as a function of the number of training images when learning on 30 object categories. The results correspond to average recognition rate over the 30 categories (each class out repeated 10 times), we run this experiment 3 times, the error bars denote $\pm$ standard deviation.

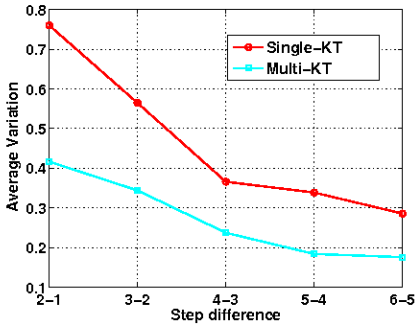


Figure 3. Norm of the differences between two $\boldsymbol{\beta}$ vectors correspondent to two subsequent steps in time. The norms are averaged both on the classes and on the splits. These results are obtained considering 10 randomly chosen classes.

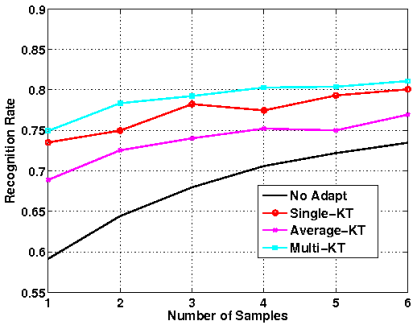


Figure 5. Classification performance as a function of the number of object training images, when learning one out of four unrelated categories. The results showed correspond to average recognition rate over the four categories, considering each class-out experiment repeated 10 times.

ber of stored models, since there is a larger probability to have stored useful prior knowledge. This intuition is confirmed by the increasing accuracy of the one-shot learning for *Multi-KT*. *Average-KT* shows a decreasing behaviour, indicating that as the prior knowledge grows, the number of unrelated classes in memory usually outnumbers the related one. The performance of *Single-KT* is more or less constant except for an evident jump in performance passing from 3 to 5 categories. Finally, Figure 4(b) shows the average classification results in case of 30 categories. It is evident here that, when learning from few samples (≤ 4), *Multi-KT* outperforms both *Average-KT* and *Single-KT*. These results, jointly with those reported in the previous section, make us conclude that *Multi-KT* is the most effective knowledge transfer algorithm, compared to *Average-KT* and *Single-KT*.

5.4. Comparison with previous work

The most famous one-shot learning algorithm in the computer vision literature is [7, 8], where the authors extract a "general knowledge" from previously learned categories. Their approach makes no assumptions on the reliability of prior knowledge, which is always considered as an average of all the known classes. To compare against this method, we repeated the four classes experiment presented in [7]. Unfortunately it was not possible to reproduce exactly their experimental setting, as the features used are no more available[3], and the algorithm was not publicly released. We opted therefore for benchmarking our results against those reported in Table 1 in [7].

We considered the classes faces, motorbikes, leopards (originally spotted-cats) and airplanes. The average recognition rate over the categories as a function of the number of object training images is shown in Figure 5. Table 1 compares the results of *Multi-KT* and *Single-KT* to that reported in [7] considering also the best one-shot result per class. This analysis confirms that our method performs better than *Single-KT*, and it obtains results comparable to [7].

6. Conclusions

We presented an SVM-based method for learning object categories from few examples. The algorithm transfers prior knowledge selecting a subset of the known classes

[3] L. Fei Fei, personal communication.

Algorithm	Error Rate (%) on 5 pos. images	Best Rec. Rate (%) on 1 pos. image	Remarks
Multi-KT	8-29	airplanes: 90.8	+6 backgr. images
Single-KT [22]	10-29	airplanes: 88.1	+6 backgr. images
[7]	8-22	faces: 82.0	

Table 1. Comparison between our *Multi-KT* algorithm, *Single-KT* [22], and the Bayesian One-Shot learning method presented in [7]. Since both *Multi-KT* and *Single-KT* are a discriminative approaches, besides the positive samples we need few background images in the training set.

and weighting them appropriately. It decides automatically from where and how much to transfer, adapting the old models to the incoming data and solving a convex optimization problem which minimizes an estimate of the generalization error. Experiments show that it outperforms both the results obtained in [22] and those produced using an average of all the previous experience. This last choice can induce negative transfer, in particular when the number of known category increases. On the contrary, when prior knowledge grows our method shows a one-shot learning behaviour. By using the features provided in [10] and making available our code[4] we are offering to the community a reproducible experimental setting that can be used in the future to test new knowledge transfer algorithms. By using several features we also showed that the behaviour of the method is not affected by the feature's choice. Future work will investigate ways to reduce the computational complexity of the algorithm for large number of known categories and analyze its asymptotical behaviour when the number of training samples increases.

7. Acknowledgments

This work was supported by the EU project DIRAC (FP6-0027787), by ....., and by the EMMA project thanks to the Hasler foundation (www.haslerstiftung.ch).

References

[1] I. Biederman. Recognition-by-components: A theory of human image understanding. *Psychological Review*, 1987. 1

[2] A. Bosch, A. Zisserman, and X. Munoz. Representing shape with a spatial pyramid kernel. In *CIVR*, 2007. 5

[3] O. Bousquet and A. Elisseeff. Stability and generalization. *JMLR*, 2:499–526, Mar 2002. 4

[4] G. Cawley. Leave-one-out cross-validation based model selection criteria for weighted LS-SVMs. In *IJCNN*, 2006. 3

[5] N. Cristianini and J. Shawe-Taylor. *An Introduction to Support Vector Machines*. Cambridge University Press, 2000. 4

[6] M. Everingham, L. V. Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The PASCAL Visual Object Classes Challenge 2008 Results. http://www.pascal-network.org/challenges/VOC/. 1

[7] L. Fei-Fei, R. Fergus, and P. Perona. A bayesian approach to unsupervised one-shot learning of object categories. In *ICCV*, 2003. 1, 2, 5, 7, 8

[8] L. Fei-Fei, R. Fergus, and P. Perona. One-shot learning of object categories. *PAMI*, 28:594–611, 2006. 2, 5, 7

[9] M. Fink. Object classification from a single example utilizing class relevance metrics. In *NIPS*, 2004. 2

[10] P. Gehler and S. Nowozin. On feature combination for multiclass object classification. In *ICCV*, 2009. 5, 8

[11] J. Gibbons. *Nonparametric Statistical Inference*. New York: Marcel Dekker, 1985. 5

[12] G. Griffin, A. Holub, and P. Perona. Caltech 256 object category dataset. Technical Report UCB/CSD-04-1366, California Institue of Technology, 2007. 1, 5

[13] D. Hush, P. Kelly, C. Scovel, and I. Steinwart. Qp algorithms with guaranteed accuracy and run time for support vector machines. *JMLR*, 2006. 4

[14] D. Lowe. Distinctive image features from scale-invariant keypoints. *IJCV*, 60(2):91–110, 2004. 5

[15] F. Orabona, C. Castellini, B. Caputo, E. Fiorilla, and G. Sandini. Model adaptation with least-squares svm for hand prosthetics. In *ICRA*, 2009. 2, 3

[16] S. J. Pan and Q. Yang. A survey on transfer learning. Technical report, Hong Kong University of Science and Technology, Hong Kong, China, November 2008. 1, 2

[17] A. Quattoni, M. Collins, and T. Darrell. Transfer learning for image classification with sparse prototype representations. *CVPR*, 2008. 2

[18] M. Rosenstein, Z. Marx, and L. P. Kaelbling. To transfer or not to transfer. In *NIPS Workshop on Transfer Learning*, 2005. 2

[19] E. Soria, J. Martin, R. Magdalena, M. Martinez, and A. Serrano. *Handbook of Research on Machine Learning Applications*, chapter L. Torrey and J. Shavlik, Transfer Learning. IGI Global, 2009. 1

[20] J. Suykens, T. V. Gestel, J. D. Brabanter, B. D. Moor, and J. Vanderwalle. *Least Squares Support Vector Machines*. World Scientific, 2002. 1, 3

[21] S. Thrun. Is learning the n-th thing any easier than learning the first? In *NIPS*, 1996. 1

[22] T. Tommasi and B. Caputo. The more you know, the less you learn: from knowledge transfer to one-shot learning of object categories. In *BMVC*, 2009. 1, 2, 3, 4, 5, 8

[23] O. Tuzel, F. Porikli, and P. Meer. Human detection via classification on riemannian manifolds. In *CVPR*, 2007. 5

[24] P. Wu and T. G. Dietterich. Improving svm accuracy by training on auxiliary data sources. In *ICML*, 2004. 2

[25] A. Zweig and D. Weinshall. Exploiting object hierarchy: Combining models from different category levels. In *ICCV*, 2007. 2

[4] ....WEBPAGE...