

Sparse Representations and Distance Learning for Attribute based Category Recognition

Grigorios Tsagkatakis¹ and Andreas Savakis²

¹ Center for Imaging Science, ² Department of Computer Engineering
Rochester Institute of Technology, Rochester, NY 14623
{gxt6260, andreas.savakis}@rit.edu

Abstract. While traditional approaches in object recognition require the specification of training examples from each class and the application of class specific classifiers, in real world situations, the immensity of the number of image classes makes this task daunting. A novel approach in object recognition is attribute based classification, where instead of training classifiers for the recognition of specific object class instances, classifiers are trained on attributes of the object images and these attributes are subsequently used for the object recognition. The attributes based paradigm offers significant advantages including the ability to train classifiers without any visual examples. We begin by discussing a scenario for object recognition on mobile devices where the attribute prediction and the attribute-to-class mapping are decoupled in order to meet the specific resource constraints of mobile systems. We next present two extensions on the attribute based classification paradigm by introducing alternative approaches in attribute prediction and attribute-to-class mapping. For the attribute prediction, we employ the recently proposed Sparse Representations Classification scheme that offers significant benefits compared to the previous SVM based approaches, such as increased accuracy and elimination of the training stage. For the attribute-to-class mapping, we employ a Distance Metric Learning algorithm that automatically infers the significance of each attribute instead of assuming uniform attribute importance. The benefits of the proposed extensions are validated through experimental results.

Keywords: Attribute Based Object Recognition, Sparse Representations Classification, Distance Metric Learning.

1 Introduction

The proliferation of camera-equipped mobile phones has generated a new set of opportunities as well as challenges for the computer vision community. One of these challenges is object recognition and image classification in resource constrained environments where processing power, available memory and bandwidth play a critical role. For example, imagine the scenario where a user captures an image with a camera-equipped smartphone and would like to learn more about the depicted object. A traditional object recognition system would either transmit the image to a server or perform some type of feature extraction and transmit the extracted features. The serv-

er would then have to perform a series of tests based on class specific classifiers, in order to identify the class of the depicted object and report back useful information.

There are two important issues regarding the feasibility of such a scheme in large scale scenarios. The first one is the underlying design assumption that a number of labeled examples are available during the training of the classifier which has to learn to predict the appropriate image class when new test examples from the same distribution are presented. Modern image classification schemes are becoming exceptional in this task, exhibiting high classification accuracy in challenging image datasets [28]. However, the traditional paradigm of training/testing examples may become too restrictive when real life classification problems are considered. In other words, collecting a number of training examples, even a small one, may not be feasible due to the sheer volume of the possible image classes. In addition, training classifiers (usually binary) may also be impractical, while the real-time application of these classifiers in a server for a large number of users will significantly degrade the performance in terms of response time.

The second issue that may lead to failure of this particular system design is more closely related to the specific case of mobile systems. Mobile systems are limited in processing capabilities, power availability and bandwidth. Transmitting raw images will quickly drain the available battery power. In addition, the transmitted information load will create congestion on the network and on the server which will directly affect user satisfaction.

In response to these challenges, the recently proposed paradigm of attribute based image classification attempts to learn attributes in place of traditional image classes. In the camera phone example shown in Figure 1, the system could identify the attributes “Has head” and “Has Arms” in order to target specific classes like human or statue. Then, the attribute “Has Skin” could be used to distinguish between a human and a statue. The various stages of the processing pipeline can be identified in Figure 1. The attribute vector generation process takes place on the mobile device while the final classification that takes into account both the attribute vector and the attributes-to-classes mapping, which is learned off-line from textual information, is performed on the server.

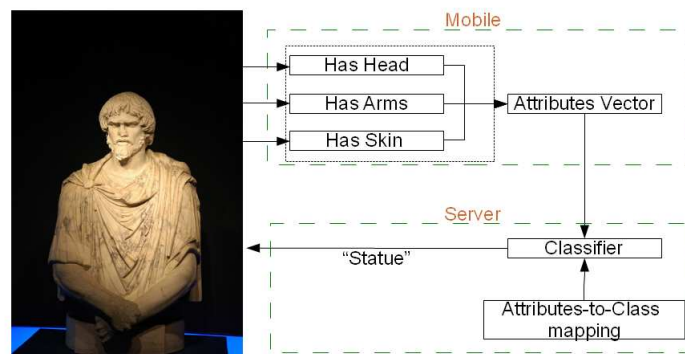


Figure 1: Attribute based image classification on a mobile system

Having attributes as an intermediate layer between image classes and image examples can provide a number of benefits. The most important gain stemming from the use of attributes is that classifiers can be trained and applied using text only information without any training images. This means that large knowledge databases can be used for the extraction of the necessary information instead of using metadata such as tags in order to automatically identify training examples. In addition, textual information is easier to store, process and transmit. The text can be used to guide an image retrieval system in order to reduce the search time for a specific query in large image databases since attributes are much smaller in dimensions and easier to handle compared to raw images. This benefit can be very significant for a system such as the one shown in Figure 1, since the server may be required to answer a larger number of queries at the same time.

As far as the mobile systems are concerned, attributes can be used in scenarios where communication bandwidth or power availability limits the amount of raw images or image descriptors that can be transmitted. In addition, the mobile system is only required to perform two processing steps, feature extraction and attribute prediction, instead of the full scale multiple classifier object recognition. Feature extraction and image classification on mobile devices, such as smartphones, has been applied in [20, 21, 27] with very promising results.

The rest of the paper is organized as follows: Previous work and the motivation of the proposed extensions are presented in Sections 2 and 3 respectively. Attribute prediction via the Sparse Representations framework is discussed in Section 4. Section 5 presents the application of Distance Metric Learning for the attribute-to-class mapping. Experimental results are presented in Section 6 and the paper concludes with a general discussion in Section 7.

2 Previous Work

Attribute based image classification is a novel paradigm in image classification where attributes are used in order to leverage the lack of training examples. There are two key scenarios where attributes have been used thus far. In the first scenario, attributes are used in order to enhance the prediction accuracy of typical classifiers when only a small number of training examples are available or the classification task is a challenging one. Examples of such cases include face verification [8], color and texture recognition [7], object detection [9] and people searching [30].

Another scenario more closely related to our work is attribute based classification where the system has to be trained without any training examples or when examples from a limited number of classes are available. In [2], Lampert et al. proposed the use of attributes for object recognition by examining the use of attributes as a midlevel layer that was used for class prediction without any training examples. In similar spirit, Farhadi et al. [1] proposed an object category recognition scheme where attribute classifiers were trained using selected features (one classifier per attribute) and the object's category was identified by applying the individual attribute classifiers on the images. The selection of features employed a L1-regularized logistic regression for the identification of class independent attribute prediction and a series of random

comparisons between class attributes and the subsequent application of linear SVMs for the final selection of the discriminative attributes. Once the relevant attributes were identified, classification was performed by selecting the class whose attributes are closest to the predicted ones.

An open issue regarding the attribute based classifier is the method by which the attributes are identified. Ideally, one would like to make this approach as unsupervised as possible. Knowledge transfer via automatic attribute identification learning was investigated in [6] where the authors used linguistic knowledge databases in order to discover the semantic link between known and unknown object classes. A similar idea was investigated in [10], where natural language processing was combined with attribute prediction in order to identify a generative model for image class recognition.

While most previous approaches utilize SVM for attribute prediction, in this work we employ the Sparse Representation Classifier (SRC). SRC was recently applied for multi-label image decomposition in [5]. The method applies the SRC framework in order to predict the labels associated with a test image by sparsely representing the label set of the test image on the label set of training examples, which is treated as the dictionary. The experimental results reported in the paper indicate the power of the SRC method for multi-label classification. Our work differs from [5] in that we apply the SRC method for attribute prediction and investigate its benefits for cross-category generalization, while [5] applies the method for traditional image based class prediction. Furthermore, in our work the predicted attributes are processed by the DML in order to identify the particular class, as opposed to the case where images of training examples are given for all classes.

3 Motivation

In this paper, we address two aspects regarding the use of attributes for image classification: attribute prediction and attribute-to-class mapping. Attribute prediction is the process where an image is presented to the system and the most prominent attributes of this image are identified. In previous works such as [1, 2], attribute specific SVM classifiers were applied in order to identify the presence or absence of a particular attribute. Although SVM is a powerful classifier, it can be expensive to train and apply during testing. In this paper, we propose the application of a novel approach in image classification termed Sparse Representations Classification method (SRC) for attribute predictions which in contrast to SVM, can be applied *without* any prior training, making it ideal for scenarios where training data is scarce and processing power limited.

The requirements for a sparse solution imposed by the SRC is intuitively appealing, since we expect to be able to use a small number of training images to represent a new test image given a specific set of attributes. Furthermore, once the sparse representation is obtained, identifying the presence or absence of a particular attribute can be rapidly evaluated by comparing the reconstruction error incurred by dictionary elements with active attribute with the error incurred by elements with inactive attribute. This approach offers higher prediction accuracy, much faster application and

exhibits higher scalability capabilities compared to the application of N successive SVM classifiers, where N is the number of attributes.

The second aspect of attribute based classification that we investigate here is the attributes-to-class mapping i.e. how to identify a particular class given a number of identified attributes. In [1], the class of a test sample is identified by looking at the attributes-to-class relationships and selecting the class whose attributes are the most similar to the predicted ones. In [2], the authors apply a Bayes classification scheme, which assumes that the presence or absence of a particular attribute is independent of the rest of the attributes. In this paper we compare the application of the Nearest Neighbor (NN) classifier with two types of distance metrics for identifying the relationship between predicted attributes and image classes. The first metric is the typical off-the-shelf Euclidean distance. In order to utilize the interdependence between different attributes, we apply a Distance Metric Learning (DML) algorithm for the discovery of the connection between image attributes and object classes.

4 Sparse Representations for Attribute Prediction

Given a signal such as a vectorized image $x \in \mathbb{R}^n$, the signal x is called k - sparse with respect to a dictionary $D \in \mathbb{R}^{n \times m}$ if $x = Ds$ where $k = \|s\|_0$ and $\|\cdot\|_0$ is the zero pseudo-norm, counting the number of non-zero elements. A linear transformation $R \in \mathbb{R}^{d \times n}$ can be applied in order to reduce the dimensionality of x from n to d , where $d \ll n$. When R is a random matrix, i.e. each element of R is drawn i.i.d. from an appropriate distribution (Gaussian, Rademacher, etc.), then the matrix R is called Random Projections (RPs) matrix [19]. Traditionally, once dimensionality reduction takes place, recovery of the original signal was not possible. However, the novel field of compressed sensing (CS) [11, 12] predicates that the original signal can be recovered from the low-dimensional representation $y = Rx = RDs$, if x is sparsely represented in some appropriate basis or dictionary D . The solution is given by the regularized ℓ_0 minimization:

$$\min \|s\|_0 \text{ subject to } y = RDs \quad (1)$$

Unfortunately, solving Equation (1) is an NP-hard problem which makes it impractical. Nevertheless, the theory of CS has shown that if the solution is sufficiently sparse, then it can be found by solving the following tractable ℓ_1 minimization:

$$\min \|s\|_1 \text{ subject to } y = RDs \quad (2)$$

A number of approaches have been proposed for solving Equation (2) such as orthogonal matching pursuit (OMP), basis pursuit (BP) and Least Absolute Shrinkage and Selection Operator (LASSO) [23] among others. When noise affects the signal or the signal is approximately sparse (the coefficients follow the power law), then the following problem called basis pursuit denoising (BPDN) can be solved:

$$\min \|s\|_1 \text{ subject to } \|y - RDs\|_2 \leq \epsilon \quad (3)$$

In this work, each new test image is sparsely represented in a dictionary by solving Equation (3). Once the sparse representation of the test image is identified we can use the identified dictionary elements for attribute prediction. Two approaches were investigated regarding the attribute prediction. In the first approach the presence or absence of an attribute is determined by measuring the reconstruction error obtained by dictionary elements that have this attribute, indicated by D^+ , versus error obtained by elements that do not have this attribute, indicated by D^- . Formally, the value of a particular attribute is set to:

$$a_i = \begin{cases} 1 & \text{if } \|y - RD^+s\|_2 \geq \|y - RD^-s\|_2 \\ 0 & \text{otherwise} \end{cases} \quad (4)$$

The advantage of this approach is that it is very easy to evaluate, since for each attribute, only the ℓ_2 error needs to be calculated. This approach is similar in spirit with [3], where person identity was established by measuring the reconstruction error with respect to examples from every individual. We term this approach as *SRC with positive-negative split* to differentiate it from the other classification approaches.

In addition to the previous approach, we also investigated an alternative approach that combined the SRC and the Nearest Neighbor Classifier (NN). In this scenario, the SRC is first applied in order to identify the dictionary elements that correspond to the sparsest approximation of the input signal. Once these elements are identified, each attribute of the test image is considered as present or absent based on the majority vote of the dictionary elements. Formally, given the set of dictionary elements S found by the SRC, the ℓ_1 nearest neighbor (L1NN) sets the attribute a_i according to the attributes of the active dictionary elements a_S as:

$$a_i = \begin{cases} 1 & \text{if } \text{mode}(a_S) == 1 \\ 0 & \text{otherwise} \end{cases} \quad (5)$$

An important question regarding the application of the SRC framework for image classification is the technique that is employed for the dictionary construction. In general, there are two approaches in designing the dictionary. The first one is to try to represent a new test image as a sparse linear combination of the full collection of training images which is treated as the dictionary. This approach has been successfully applied in various computer vision tasks such as face recognition [3, 4].

The second approach tries to identify a small number of elements that are adequate for representing the training examples. In the majority of recognition systems, a generative visual vocabulary is constructed by applying the k -means clustering algorithms on the low level features e.g. [24]. However, issues like the lack of supervision and the explicit definition of the number of neighbors can hinder the recognition accuracy. More recent methods like the K-SVD [25] and supervised dictionary learning [26] are more focused towards the sparse representation framework by trying to generate a dictionary so that the training examples have a sparse representation.

In this work, we used the entire collection of training examples as a dictionary based on two assumptions. The first assumption is that, given the possible disassociation between training and testing set, selecting a small number of elements for the dictionary, via a generative dictionary construction method, may provide good results for the training set but poor results for the testing set. In addition, if a discriminative

approach is followed, then the SRC has to be applied for the prediction of every individual attribute independently which will severely affect the processing time for each testing example.

The second assumption is more closely related to the setup we are considering. If all the available training examples are used as dictionary elements, then we remove the requirement for training. The elimination of a training stage may be of significance in scenarios where incremental learning of the attributes is required or fast application of the attribute prediction step.

5 Distance Metric Learning for Attribute-to-Class mapping

Given a binary attribute vector indicating the presence or absence of specific attributes in a test image, classification is performed by selecting the class whose attribute vectors are *closest* to the one of the test images. In this scenario we assume that for each class, a number of active attributes are identified that are typical for this class but the images used for the attribute identification are not available to the classifier, in contrast to traditional image based systems.

The link between attribute vectors and class, i.e. how to infer the appropriate class given the attribute vector, is critical. This link is generally related to the use of a *lexicon*. By *lexicon* we mean a list of textual description for each class. Ideally, each class would correspond to a specific set of attributes, e.g. the dog class is described by “tail”, “head”, “fury” etc. However, unless such a lexicon is explicitly defined, we cannot expect to have such a clear and unambiguous description of the classes. A more realistic scenario is one where, for each class, a list of different attribute vectors is provided by an unsupervised information retrieval system. In this scenario, some of the retrieved descriptions may also contain “face” and “arm” because of images where the dog is portrayed next to his owner or “door” and “furniture” because the images show dogs in indoor settings. The goal of the attribute-to-class mapping is to infer the correct class given a number of possible attribute combinations.

An important question regarding the attribute-to-class mapping is the type of similarity metric i.e. how to measure the distance between two attribute vectors. Typically, the distance between two vectors is measured using off-the-shelf distances like the Euclidean or the Hamming distance. However, recent approaches have shown that using a distance metric learned from the available data can significantly improve the classification results. In supervised Distance Metric Learning (DML), the objective is to learn a new distance metric that will satisfy the pairwise constraints imposed by class label information. Formally, the distance between two data points x and $y \in \mathbb{R}^n$ is given by the

$$d_G(x, y) = \|x - y\|_G^2 = (x - y)^T G (x - y) \quad (6)$$

where $G \in \mathbb{R}^{n \times n}$ is a Mahalanobis-like distance. The matrix G is required to be positive semidefinite, since this property guarantees that the new distance will satisfy the requirements for a metric i.e. non-negativity, symmetry, and triangle inequality.

In this paper, we utilize a recently proposed method for local DML called Information Theoretic Metric Learning (ITML) [16]. The goal of the ITML is to minimize the

“closeness” between the Mahalanobis distance matrix G and a given Mahalanobis distance matrix G_0 while keeping the intraclass distance smaller than the interclass distance. In order to measure the “closeness” between the two distance matrices, G and G_0 , the ITML assumes that each distance matrix corresponds to the typical Mahalanobis distance of two unknown multivariate Gaussian distributions given by $p(x; G_0) = \frac{1}{Z} \exp(-\frac{1}{2} d_G(x, \mu))$, where μ is the mean and Z is a normalization constant. Then, the Kullback–Leibler (KL) divergence is employed as a robust metric of the correspondence between the two Gaussian distributions. We apply the DML framework in order to learn a distance that will bring attribute vectors from similar classes closer than attribute vectors from different classes. Formally, given two attribute vectors x and y we solve the following minimization problem:

$$\min_G KL(p(x; G_0) \parallel p(x; G)) \quad (7)$$

subject to the constrains

$$\begin{aligned} d_G(x, y) &\leq l \text{ if } class(x) = class(y) \\ d_G(x, y) &\geq u \text{ if } class(x) \neq class(y) \end{aligned} \quad (8)$$

ITML is a recently proposed approach that offers significant benefits including fast training, since it does not require the expensive eigen-decomposition and fast application to new examples [29]. Once the attribute vector is identified by the classifier, the mapping of the attribute vector to a class is performed by measuring the distance between the newly identified attribute vector and the training attribute vectors. In this stage there are various approaches on the choice of distance. In this work, we assign the class by comparing the mean distance between the test attribute vector and examples from a single class and selecting the class with the minimum average distance. Formally, the class of a new attribute vector x is given by

$$class(x) = argmin_{c_i \in C} \{mean_{y_{ij} \in c_i} d_G(x, y_{ij})\} \quad (9)$$

where $C = \{c_1, \dots, c_m\}$ is the collection of classes and y_{ij} is the j^{th} example from the i^{th} class of the training set.

6 Experimental Results

To validate the proposed extensions to the attribute based classification, we use the recently developed dataset by Farhadi et al. [1], where a large collection of images were annotated from a list of 64 attributes by Amazon Turk annotators. These attributes include the presence of particular image parts such as “head”, “ear”, “wing”, “windows” etc, overall shape such as “2D boxy”, “round”, “vertical cylinder” etc, and material attributes such as “feathers”, “plastic”, “metal” etc. The dataset consists of two parts. The first part, called a-Pascal, used the images from the PASCAL VOC 2008 dataset. This dataset consists of images from twenty classes and each class is represented by 150 to 1000 images per class. This dataset was divided in two sections

were 6340 are used for training and 6355 for testing. In order to test the ability to generalize the attribute prediction for the classification of images from unseen class, a second dataset called a-Yahoo, was utilized. The a-Yahoo dataset consists of 2644 images from 12 classes that are different from the classes of the a-Pascal.

For each image, the bounding box of each object was first determined and the attributes corresponding to the object within the bounding box were identified. In order to represent each image, the same process as in [1] was employed. More specifically, for each image a number of base features were extracted corresponding to color, texture, visual parts and edges. Texture descriptors were extracted for each pixel and k -means was applied to quantize the descriptors to 256 clusters. The HOG spatial pyramid descriptors quantized to 1000 k -means clusters were used for visual words generation and the Canny edge detector was employed for edge descriptions, quantized to 8 unsigned bins. Color information was represented by quantized color descriptors. These descriptors were applied in a grid of three vertical and two horizontal blocks in order to generate the overall 9751-dimensional feature representation of each image.

To decrease the memory and time required for training and testing, the Random Projections method [19] was applied. This reduced the dimensionality of the 9751-dimensional vector to 1000-dimensional vector. The RP matrix was generated by drawing i.i.d sample for a Rademacher distribution. This type of dimensionality reduction is natural for the SRC [3] and has minimal effects on the performance of linear SVM as it was shown in [18].

6.1 Attribute Prediction

The first set of experiments involves the prediction of the attributes for a specific image. For this experimental setup, we measured the performance in attribute prediction when training and testing examples come from the same set (within-category) and attribute prediction when training and testing sets are disjoint (cross-category). For the within-category attribute prediction, the a-Pascal dataset was used for both training and testing while for the cross-category prediction, the a-Pascal was used for training and the a-Yahoo for testing.

We tested three approaches in attribute learning. The first one is linear SVM, similar to [1], where a separate SVM classifier was trained on each attribute. The second one is the *SRC with positive-negative split* as described in Section 4. The OMP algorithm was used for the SRC which is part of the SparseLab [22]. The third one is the L1NN as described in Section 4.

Table 1 presents the classification error for within-category recognition and cross-category recognition using three metrics, the Hamming Loss, the F1 score and the Mean Accuracy. We observe that in both scenarios, the SRC achieves the best results in all three error metrics. The largest increase in prediction accuracy is observed in the cross-category scenario which is the main focus of the attribute based image classification. We note again that the SRC was not trained on a particular set of attributes in neither scenario.

We also investigated the role of each individual attribute with respect to the prediction accuracy. The results are shown in Figure 2 for the within category prediction

and in Figure 3 for the cross-category prediction (the y-axis corresponds to the mean prediction accuracy and is omitted for exposition purposes).

Table 1: Within category attribute prediction

Method		SVM	L1-NN	SRC
Within-category	Hamming	0.117138	0.111046	0.104265
	F1 score	0.035928	0.024183	0.169455
	Accuracy	88.89	88.29	89.57

Table 2: Cross category attribute prediction

Method		SVM	L1-NN	SRC
Cross-category	Hamming	12.68	10.65	9.79
	F1 score	0.023934	0.021711	0.101427
	Accuracy	87.32	89.35	90.21

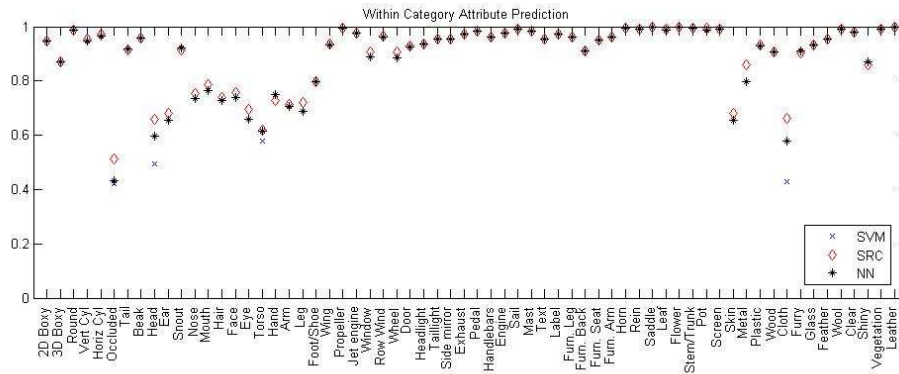


Figure 2: Individual attribute prediction on the a-Pascal dataset

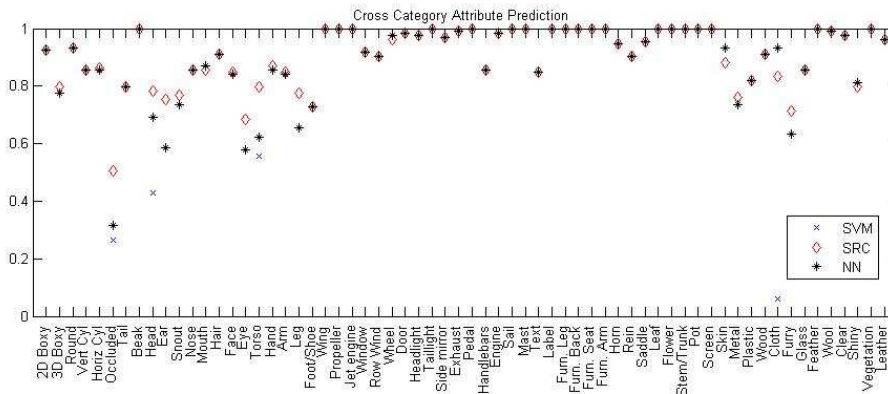


Figure 3: Individual attribute prediction on the a-Yahoo dataset

We can make two observations regarding the contribution of each attribute. First, some attributes are more important than others. For example the attribute “occluded” is very difficult to predict compared to attributes like “wing” and “sail”. This is expected, since, intuitively, the ability to identify a specific attribute is related to its ubiquitousness. Better defined attributes are easier to identify compared to more fuzzy ones.

The second observation is that attributes exhibit similar behavior in both the within-category prediction and the cross-category prediction. In other words, attributes that are easy to predict when the classes are known, remain easy to predict even if the classes are not known. For example, attributes like “leaf”, “flower” and “screen” achieve high prediction accuracy in both within and cross category prediction, in contrast to attributes like “occluded” and “cloth” which are difficult to predict in both cases. This observation further supports the argument that attributes can be reliably used for transfer learning.

6.2 Attribute-to-Class Mapping

In this section we investigate the performance of the attribute-to-class mapping. The main issue in the mapping is the size of the lexicon, i.e. given a number of classes, how many examples are necessary in order to identify the significant attributes of each class and infer the appropriate class corresponding to each attribute vector. Figures 4 and 5 present the results for the a-Pascal and the a-Yahoo dataset respectively. The a-Pascal dataset is divided into training set and testing set. The results in Figure 4 correspond to the mean accuracy obtained given a specific number of training examples per class. The a-Yahoo is significantly smaller and the classes are not balanced, thus the results in Figure 5 correspond to the mean accuracy obtained on the given number of training examples from all 12 classes.

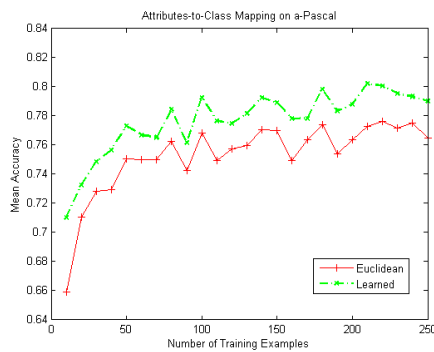


Figure 4: Attribute based class prediction on a-Pascal

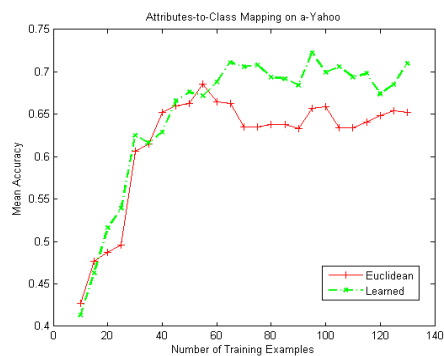


Figure 5: Attribute based class prediction on a-Yahoo

We can make two observations based on these results. Regarding the recognition accuracy, we observe that applying a learned distance outperforms an off-the-shelf distance in both cases. However, we see that although there is a general tendency to achieve better results with more examples, the corresponding curves are not smooth.

The lack of smoothness is attributed to the fact that many classes share the same attributes. This makes the process of identifying the appropriate class difficult. We expect that given either more discriminative or a larger pool of attributes, the general tendency would be more evident.

The last observation brings in focus the process by which the attributes are assigned to each class. Attribute assignment without taking into account the cases where attributes from different classes are the same, can significantly deteriorate the overall performance. This suggests the need for either a larger collection of attributes, which will make class identification easier, or a discriminative process in attribute selection.

6.3 Learning from purely textual information

In the last set of experiments we present the ability of the system to identify new classes from purely textual descriptions. The description of a class is provided as an attribute vector and the classification is achieved by selecting the most similar class based on the identified attributes. In order to evaluate the performance of the system in this challenging task, the a-Pascal training set was used to train attribute based SVM classifiers. When a new image from the a-Yahoo dataset was presented to the SVM, the L1NN and the SRC were applied in order to identify the attribute vector. The identified attribute vector was then mapped to a class based on examples from the a-Yahoo dataset. The mean classification accuracy is presented in Table 3.

Table 3: Cross Category Attribute based Class Recognition

	Euclidean	Learned
L1NN	8.67	13.15
SVM	10.66	13.15
SRC	11.56	15.87

Regarding the classification algorithm, we see that the *SRC with positive-negative split* outperforms both the L1NN and the SVM, using either the Euclidean or the learned distance. This result is especially important, since no training stage was applied for the SRC classification. As for the distance metric used for the attribute-to-class mapping, we observe that using a learned distance can provide significant benefits in terms of recognition accuracy. We note that the results presented in Table 3 are obtained using only the 64 attributes, whereas the results obtained in [1] included 1000 additional discriminative attributes generated by a random comparison process.

7 Discussion

Attribute based image classification is a recently proposed paradigm in object recognition that could support the challenging task of object recognition in resource constrained environments such as mobile devices. Under this paradigm, objects are de-

scribed by vectors that indicate the presence of particular attributes. The pipeline of the attribute based classification consists of two parts. First, given a new image the corresponding attributes are identified. In this work we propose the application of the Sparse Representation Classification framework in place of the traditional attribute specific SVM. This new framework achieves higher accuracy without any prior training. Once the attributes are identified, the mapping to a class is based solely on textual information without the need of visual examples. We propose the use of Distance Metric Learning in order to identify the importance of each attribute with respect to each class.

Considering the overall classification accuracy of our system, we maintain that the proposed system achieves better performance compared to previous approaches. Nevertheless, the recognition rates are lower compared to the rates achieved by classification schemes trained with many visual examples from each class. One reason for the lower accuracy is the limited number and overlapping attributes used for class prediction. The assumption that the presence or absence of a particular attribute is independent of the rest of the attributes could be a factor that limits the recognition capacity of the system. Compared to image based classification, the attribute based classification scheme can perform object recognition from purely textual information without any visual examples. Semantic grouping of attributes and structured recognition may be considered to increase recognition rates.

References

1. Farhadi, A. Endres, I. Hoiem, D., Forsyth, D.: Describing objects by their attributes. In: CVPR (2009)
2. Lampert, C.H., Nickisch, H., Harmeling, S.: Learning to detect unseen object classes by between-class attribute transfer. In: CVPR (2009)
3. Wright, J., Yang, A.Y., Ganesh, A., Sastry, S.S., Ma, Y.: Robust Face Recognition via Sparse Representation. In: IEEE Trans. PAMI 31(2), 210-227 (2009)
4. Huang, J., Huang, X., Metaxas, D.: Simultaneous Image Transformation and Sparse Representation Recovery. In: CVPR (2008)
5. Li, T., Mei, T., Yan, S., Kweon, I.S., Lee, C.: Contextual decomposition of multi-label images. In: CVPR 2270-2277 (2009)
6. Rohrbach, M., Stark, M., Szarvas, G., Schiele, B., Gurevych, I.: What Helps Where – And Why? Semantic Relatedness for Knowledge Transfer. In: CVPR (2010)
7. Ferrari, V., Zisserman, A.: Learning visual attributes. In: NIPS (2007)
8. Kumar, N., Berg, A.C., Belhumeur, P.N., Nayar, S.K.: Attribute and Simile Classifiers for Face Verification. In: ICCV (2009)
9. Wang G., Forsyth D.: Joint learning of visual attributes, object classes and visual saliency. In: ICCV (2009)
10. Wang, J., Markert, K., Vergham, M., Learning models for object recognition from natural language descriptions. In: BMVC (2009)
11. Candes, E.J., Romberg, J., Tao, T.: Robust uncertainty principles: exact signal reconstruction from highly incomplete frequency information. In: IEEE Trans. Info. Theory 52(2): 489–50, (2006)
12. Donoho, D.L.: Compressed sensing. In: IEEE Trans. Info. Theory 52(4): 1289–1306 (2006)
13. Xing, E., Ng, A., Jordan, M., Russell, S.: Distance metric learning with application to clustering with side-information. In: Adv. NIPS (2003)

14. Goldberger, J., Roweis, S., Hinton, G., Salakhutdinov, R.: Neighborhood component analysis. In: Adv. NIPS (2004)
15. Globerson, A., Roweis, S.: Metric learning by collapsing classes. In: Adv. NIPS (2006)
16. Davis, J.V., Kulis, B., Jain, P., Sra, S., Dhillon, I.S.: Information-Theoretic Metric Learning. In: ICML (2007)
17. Weinberger, K. Q., Saul, L.K.: Distance metric learning for large margin nearest neighbor classification. In Journal of Machine Learning Research 10: 207-244 (2009)
18. R. Calderbank, S. Jafarpour, and R. Schapire. Compressed learning: Universal sparse dimensionality reduction and learning in the measurement domain. Preprint, (2010)
19. Baraniuk, R., Davenport, M., DeVore, R., Wakin, M.A: Simple Proof of the Restricted Isometry Property for random matrices. In: Constr. Approx. 28(3): 253-263 (2008)
20. Ta, D.N., Chen, W.C., Gelfand, N., Pulli, K.: SURFTrac: Efficient tracking and continuous object recognition using local feature descriptors. In: CVPR (2009)
21. Wagner, D., Reitmayr, G., Mulloni, A., Drummond, T., Schmalstieg, D.: Pose tracking from natural features on mobile phones. In: ISMAR (2008)
22. D. Donoho, "Sparselab." Available: <http://sparselab.stanford.edu/>. Retrieved 3/2010
23. Yang, A., Ganesh, A., Sastry, S., Ma, Y.: Fast L1-Minimization Algorithms and An Application in Robust Face Recognition: A Review. In: University of California at Berkeley Technical report UCB/EECS-2010-13 (2010)
24. Lazebnik, S. and Schmid, C. and Ponce, J.: Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In: IEEE CVPR (2006)
25. Aharon, M., Elad, M., Bruckstein, A.M.: The K-SVD: An algorithm for designing of over-complete dictionaries for sparse representations. In: IEEE Trans. SP 54(11): 4311 - 4322 (2006)
26. Mairal, J., Bach, F., Ponce, J., Sapiro, G., Zisserman, A.: Supervised dictionary learning. In: Adv. NIPS 21 (2009)
27. Wagner, D., Schmalstieg, D., Bischof, H.: Multiple target detection and tracking with guaranteed framerates on mobile phones, IEEE ISMAR (2009)
28. Everingham, M., Van Gool, L., Williams, C. K. I., Winn, J., Zisserman, A.: The Pascal Visual Object Classes (VOC) Challenge. In: IJCV 88(2): 303-338 (2010)
29. Jain, P., Kulis, B., Dhillon, I., Grauman, K.: Online Metric Learning and Fast Similarity Search. In: Adv. NIPS (2008)
30. Vaquero, D.A., Feris, R.S., Tran, D., Brown, L., Hampapur, A., Turk, M.: Attribute-Based People Search in Surveillance Environments. In: IEEE WACV (2009)