COSTA: Co-Occurrence Statistics for Zero-Shot Classification

Mensink, T.; Gavves, E.; Snoek, C.G.M.

DOI
10.1109/CVPR.2014.313

Publication date
2014

Document Version
Author accepted manuscript

Published in
Proceedings: 2014 IEEE Conference on Computer Vision and Pattern Recognition: 23-28 June 2014, Columbus, Ohio

COSTA: Co-Occurrence Statistics for Zero-Shot Classification

Thomas Mensink Efstratios Gavves Cees G.M. Snoek
ISLA, Informatics Institute, University of Amsterdam

Abstract

In this paper we aim for zero-shot classification, that is visual recognition of an unseen class by using knowledge transfer from known classes. Our main contribution is COSTA, which exploits co-occurrences of visual concepts in images for knowledge transfer. These inter-dependencies arise naturally between concepts, and are easy to obtain from existing annotations or web-search hit counts. We estimate a classifier for a new label, as a weighted combination of related classes, using the co-occurrences to define the weight. We propose various metrics to leverage these co-occurrences, and a regression model for learning a weight for each related class. We also show that our zero-shot classifiers can serve as priors for few-shot learning. Experiments on three multi-labeled datasets reveal that our proposed zero-shot methods, are approaching and occasionally outperforming fully supervised SVMs. We conclude that co-occurrence statistics suffice for zero-shot classification.

1. Introduction

Zero-shot classification aims to reveal the relevant class of an image, in the case where no visual examples of that class are provided during training [12, 16, 20]. In the absence of direct annotated data, visual classes should be described and classified indirectly. This indirect classification usually takes place in two stages. First, the visual appearance of object classes is described using semantic properties, such as attributes [12, 28] or class hierarchies [16, 20]. Second, a transfer scheme has to be provided at test time for the new (unseen) class, *e.g.*, an attributes-to-class mapping, or its position in the hierarchy.

In this paper we introduce COSTA, using the co-occurrence statistics of visual concepts for transfer learning. First, we use a set of known labels as knowledge data, without requiring attribute annotation or a specific hierarchy. Second, our transfer scheme relies on co-occurrence statistics between the new class and the known labels. These are easy to obtain, *e.g.*, by active learning, web engines or user-provided image tags. Our approach is illustrated in Figure 1.

COSTA leverages, by design, the bias of natural co-occurrences of visual concepts. These emerge naturally in complex images, when multiple concepts appear together in images. The underlying hypothesis is that concept-to-concept inter-dependencies reveal a significant part of the latent image semantics. This bears three important advantages as compared to attribute-based classification [12].

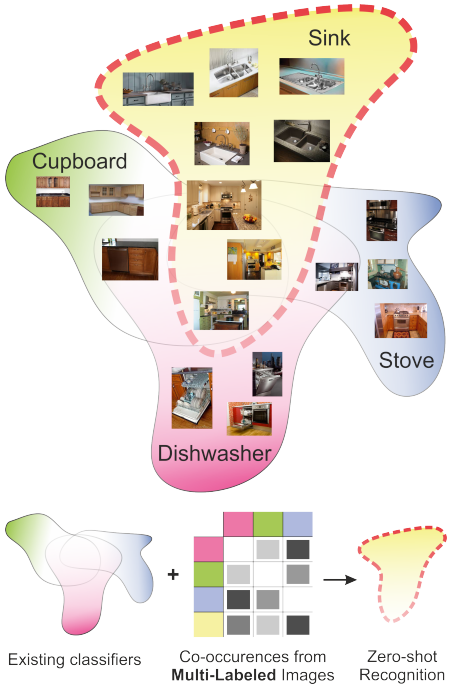

Figure 1. Illustration of COSTA, the classifier for an *unseen label* is estimated using a weighted combination of existing classifiers and their co-occurrence statistics.

First, many concepts can be described as an open set of concept-to-concept inter-dependencies. For example a *chair* is probably better described by contextual cues that are easier to recognize, such as *indoors*, *table* or *desk*, rather than by its composing parts that we may not even be able to define [7, 14]. This contrasts to the restricting assumption of attribute-based classifiers where a class is a closed set of object specific semantic attributes.

Second, the mapping between unknown concepts and the known labels requires only computing their respective co-occurrences. Hence, we avoid the high-level mappings from

attributes to classes, which are challenging to automate [21] and often need to be provided by specialists [12, 28]. Co-occurrence statistics are relatively straightforward to obtain from external sources.

Third, in our framework all visual concepts are treated equally, objects, scene descriptions and classes are all *labels*. There is no need to distinguish between *classes*, to be observed, and *attributes*, to describe classes. Therefor, COSTA inherently allows for zero-shot recognition in *multi-labeled* datasets, for which we are the first to provide a principled method. Moreover, since we make no assumptions on the nature of concepts, we may as well combine object classes, attributes, scene descriptions and any other textual image description as labels.

In fact, we observe that within a natural image there are usually many relevant concepts [29], although just a few of them are annotated [5]. As a result, even if we have the image features for learning an unknown concept, we cannot actually facilitate the learning in the absence of the actual labels. Exploiting, however, inter-concept dependencies unlocks the learning of such unknown concepts. This is particularly useful for concepts whose appearance cannot be easily learned, either due to the small object size, large intra-class variation, or the scarcity of annotations.

We summarize our novelties as follows. First, in this paper we introduce the use of co-occurrence statistics for zero-shot recognition, which is applicable for *multi-label* zero-shot classification. Second, we show that for certain types of concepts it is better to learn an indirect model, based on concept-concept relationships, rather than directly using visual examples. Third, we demonstrate that co-occurrence statistics could also be obtained from web search engines. This allows for effortless zero-shot classification, albeit with an expected decrease of performance. Forth, we show that our zero-shot recognition model based on concept co-occurrences can serve as a prior in a few-shot learning setting, which significantly improves performance when just a few positive instances are available.

2. Related work

In this section we discuss some of the most relevant work to zero-shot learning, few-shot learning, and using co-occurrence statistics.

Zero-shot learning. Due to the ever increasing number of available images and categories to be recognized it becomes infeasible to label images for each possible class. In their pioneering work [12] Lampert *et al*. propose to use an attribute-based representation to capture semantic properties of an image, for multi-class zero-shot classification. This is an extreme case of transfer learning where for a new class no training instances are available, only a description of a class in terms of attributes. For learning animal classes, for example, these attributes could consists of *has fur* and *eats fish*. Given a set of trained attribute classifiers, an image is classified to a class by comparing its attribute predictions to a set of predefined attribute-to-class mappings.

The work has generated momentum for zero-shot classification, and the attribute-based representation has developed in several ways. For example, by including learned non-semantic attributes for better discrimination [19, 25], or to learn an attribute embedding specific for zero-shot prediction [1]. Instead of relying on a predefined mapping, in [21], linguistic knowledge databases and web search hit counts are used to automatically obtain the attribute-to-class mapping. They obtained the best automated results by using the *hit counts* of the Yahoo search engine, which we will also use in our experiments.

Most of these works exploit expensive, expert-driven, annotations to obtain semantic attributes and/or rely on difficult to obtain class-to-attribute mappings. Furthermore, they all focus on *multi-class* classification of images that contain a single main object. In this paper we also focus on zero-shot classification, but our method is suitable for *multi-label* image classification. To describe novel concepts, we exploit label-to-label relationships, which are easy to obtain under the assumption that visual concepts have unique co-occurring patterns in images.

Few-shot learning. In few-shot learning it is assumed that besides a few positive instances also example images from (loosely) related classes are provided [13]. The goal is to exploit the related classes to improve the classification accuracy. Hierarchies of objects and classes are conceptually appealing for such kind of knowledge transfer. For example, the WordNet hierarchy has been used to obtain zero-shot priors for large scale image classification [16, 20]. In [23] hierarchies are used to transfer knowledge from well represented classes to related classes with just a few examples. And in [2] the authors transfer valuable features from already known classes for describing a novel class using feature adaptation. In our work, we do not impose any hierarchy on the objects, instead we use the co-occurrences of visual concepts to transfer knowledge to new classes.

In [20], various knowledge-transfer methodologies are evaluated for few-shot and zero-shot recognition. The authors conclude that knowledge transfer has little added value when ample training images are available for all classes. In contrast, transfer learning was found to be effective in a zero-shot and few-shot classification setting. In the current work we *do not* aim to improve upon methods that focus on few-shot learning, instead we illustrate that our zero-shot prediction model serves as a reasonable *prior* for few-shot classification. To do so, we rely on the the weighted least-square support vector machines for learning from few examples [26].

Co-occurrence statistics. Co-occurrences have been repeatedly considered for capturing higher order relationships between classes, concepts and labels. They have been used to improve image segmentation [11], object detection [3, 22], to reason about what to expect in a scene [4, 14], and for image and attribute classification [8, 15]. In general, we notice that co-occurrences of objects, labels and textures have been recognized as a strong clue for label and attribute prediction. However, we are not aware of any works that use label co-occurrences to enable zero-shot classification.

3. Zero-shot multi-label classification

In this section we define our zero-shot framework using label co-occurrences. We first introduce our co-occurrence based classification method. Second, in Section 3.2 we describe a regression interpretation to obtain zero-shot classifier. Third, in Section 3.3 we illustrate the use of the zero-shot model as prior for few-shot classification.

3.1. Co-occurrence based classification

Our goal is to estimate a classification function for an *unseen label* l, using a set of existing classifiers. We assume that we have a set of linear classifiers, trained on a collection of annotated images with k labels. These classifiers could be obtained from binary SVMs or logistic regression, and are represented by their weight vectors $\boldsymbol{w}_k \in \mathbb{R}^{d\times 1}$. We assume, without loss of generality, that the weight vectors are augmented such that the biases are included.

We propose to estimate the weight vector $\hat{\boldsymbol{w}}_l$ to classify the unseen label l, as

$$\hat{\boldsymbol{w}}_l = \sum_k \boldsymbol{w}_k \, s_{lk}, \tag{1}$$

where s_{lk} represents a measure of similarity between the known label k and the unseen label l. In this paper, we base these similarities on the co-occurrence statistics between the new label and existing labels.

Co-occurrence similarities. We explore different similarity measures based on the co-occurrence of two labels. Let c_{ij} denote the total number of images for which label i and label j are relevant according to an auxiliary resource, for example the ground-truth labelling, a web search engine or a user provided input in the case of active learning. Also, c_i denotes the total number of images depicting label i, and m denotes the total number of labels. The similarities we explore are:

- Normalized co-occurrences,

$$s^{\mathrm{n}}_{ij} = \frac{c_{ij}}{c_i}, \tag{2}$$

  where the similarity is directly proportional to the number of co-occurrences.

- Binarized co-occurrences, motivated by the binarized class-to-attributes mappings used in attribute-based zero shot classification[12, 28]:

$$s^{\mathrm{b}}_{ij} = [\![c_{ij} \geq t]\!], \tag{3}$$

  where t is a global threshold set as $t = \frac{1}{m^2}\sum_{i,j} c_{ij}$.

- Burstiness corrected co-occurrences. Burstiness is the phenomena that the frequency of an observation is significantly larger than a statistically independent model would predict. The square-root function is used in image retrieval and classification to correct for burstiness of visual features [10, 24]. Similarly, we aim to reduce the burstiness in labellings, and use:

$$s^{\mathrm{s}}_{ij} = \sqrt{c_{ij}}, \tag{4}$$

- The Dice's coefficient,

$$s^{\mathrm{d}}_{ij} = \frac{c_{ij}}{c_i + c_j}, \tag{5}$$

  is a measure used in many Natural Language Processing systems. It aims to estimate the semantic relatedness between two terms, based on hit counts from web search engines.

Defining a concept by what it is not. The similarities defined above, are based only on positive co-occurrences, *i.e.*, how often two labels are relevant for an image. However, knowing what is *not* related to the concept might be a very informative clue about the visual scope of a concept, which is also shown in an image retrieval setting [9].

In addition to the positive co-occurrences, denoted by c^{++}_{ij}, we also use the other possible co-occurrence relations: the presence of label i with the absence of label j, the absence of label i with the presence of label j, and the absence of both labels, denoted by c^{+-}_{ij},c^{-+}_{ij}, and c^{--}_{ij} respectively. For each of these definitions of co-occurrence we use the similarity measures defined above.

Using the positive and negative co-occurrences, the weight vector $\boldsymbol{w}_l$ of an unknown label can be estimated as:

$$\hat{\boldsymbol{w}}_l = \sum_k \boldsymbol{w}_k \, s^{++}_{lk} - \boldsymbol{w}_k \, s^{+-}_{lk} - \boldsymbol{w}_k \, s^{-+}_{lk} + \boldsymbol{w}_k \, s^{--}_{lk}, \tag{6}$$

Estimate co-occurrence from web data. So far, we have not discussed how to obtain the co-occurrence statistics required for our zero-shot recognition framework. To show the potential of our framework, in most of the experiments we estimate label co-occurrences from the ground-truth labelling of our image datasets. Alternatively, the co-occurrence statistics could be estimated from large text corpora, *e.g.*, Wordnet or Wikipedia, or internet search engines, such as Yahoo, Google and Bing.

Following [21], we use the *hit counts* estimations of the Yahoo web search, Yahoo image search and Flickr-tag search engines to estimate to co-occurrences. We estimate the similarities defined above, using c_{ij}^{HC} denoting the hit count for a query consisting of label i and j. The positive and negative co-occurrences are estimated by using the hit counts c_i^{HC} of the individual labels i, and an estimate of the total number of images $c_{\text{total}}^{\text{HC}} = \sum_i c_i$.

3.2. Regression to estimate classifiers

To improve the classifier for the unseen label l, we propose to learn a weighted version of Eq. (1), given by:

$$\hat{\boldsymbol{w}}_l = \sum_k a_k \, \boldsymbol{w}_k \, s_{lk}, \tag{7}$$

where a_k is a weight for classifier k, which is independent from the unseen label l.

Ideally, we set the classifier weights $\boldsymbol{a} \in \mathbb{R}^{k \times 1}$ such that, the estimated weight vector $\hat{\boldsymbol{w}}_l$ equals the *ideal* weight vector $\boldsymbol{w}_l$, which would have been be obtained by learning on the visual data with annotations available. This could be seen as a regression problem, where $\boldsymbol{a}$ is set to regress $\hat{\boldsymbol{w}}$ towards $\boldsymbol{w}$. However, since we aim for zero-shot classification, we do not have access to the unseen labels l, nor the ideal weight vectors $\boldsymbol{w}_l$ at train time. Therefore we use the known labels k in a leave-one-out setting for learning.

We minimize the following regression squared-loss:

$$L_{\text{reg}} = \sum_i \|\boldsymbol{w}_i - \sum_k a_k \boldsymbol{w}_k \, s_{ik}\|_2^2, \tag{8}$$

$$= \sum_i \sum_d \left(w_{id} - \boldsymbol{a}^\top \boldsymbol{v}_{id}\right)^2, \tag{9}$$

where index i and k both run over the known labels, and $s_{ii} = 0$. The vector $\boldsymbol{v}_{id}$ contains the k co-occurrence weighted weight vectors $v_{idk} = s_{ik} \, w_{kd}$.

Note that the loss is formulated over train classes and not over train images. Moreover, Eq. (9) shows that $\boldsymbol{a}$ can be obtained in closed-form using ridge-regression. In practice we observe that regularization is not needed for good performance, the dimensionality of $\boldsymbol{a}$ is much smaller than the number of training instances (k vs. $d\,k$).

3.3. Zero-shot prior for few-shot prediction

In a few-shot classification setting a few, *e.g.*, up to 8, positive images are provided per label to learn a classifier. In such a setting a strong prior could benefit the performance by guiding the SVM classifier. In this section we consider a simple model adaptation method where the zero-shot model acts as a prior for the few-shot classifier.

In the case that we employ linear SVM classifiers with squared hinge-loss, the objective to minimize becomes:

$$L_{\text{few}} = \frac{C}{2} \sum_i \left[1 - y_{il} \, \boldsymbol{w}_l^\top \boldsymbol{x}_i\right]_+^2 + \frac{1}{2} \|\boldsymbol{w}_l - \beta \, \boldsymbol{w}_z\|_2^2, \tag{10}$$

where $\boldsymbol{w}_z$ is the prior obtained from the zero-shot model, and $\beta \in (0, 1)$ is a scaling parameter to control the degree to which the label classifier should be similar to the zero-shot model. It can be shown that the optimal solution for $\boldsymbol{w}_l$, for a specific value of C, is given by [18, 27]:

$$\hat{\boldsymbol{w}}_l = \boldsymbol{w}_g + \beta \boldsymbol{w}_z, \tag{11}$$

where $\boldsymbol{w}_g$ is the weight vector obtained from optimizing the standard SVM formulation, *i.e.*, using Eq. (10) with $\beta = 0$. We will use $\boldsymbol{w}_g$ also as a baseline few-shot classifier. Note that the optimal parameter C could differ when including the prior. In our experiments, we first determine the optimal value of C for $\boldsymbol{w}_g$ using cross-validation, then we obtain $\boldsymbol{w}_l$ by using $\beta = 1$.

4. Experiments

In this section we experimentally validate our models for zero-shot and few-shot image labelling using co-occurrence statistics. Since we are interested in a zero-shot classification setting, for most experiments we split the labels of the datasets into two *disjoint* sets: the *known* classes and the *unseen* classes. The true zero-shot classifiers use labels from the known classes *only*. Since we are not aware of any methods that do zero-shot recognition on multi-label data sets nor using multi-label co-occurrence statistics, we compare with the binarized version of co-occurences, the closest to what attributes could be for multi-labeled datasets. Furthermore, for all datasets we report results obtained in a fully supervised setting, where the SVM classifiers use the training labels from *all* classes, be it known or unseen.

4.1. Experimental set-up and data sets

Image features. For all our experiments we use the Fisher Vector (FV) image representation [24]. Per image a single FV $\boldsymbol{x}$ is extracted, and we follow a common pipeline and use: (i) SIFT descriptors, projected with PCA to 96-dimensions; (ii) Mixture-of-Gaussian codebook with 16 components; (iii) power-normalization and ℓ_2 normalization. The final FV is only 3K dimensional, despite the compactness its performance is still competitive.

Implementation. For all experiments where SVM classifiers are used, we train linear SVMs [6] and employ two-fold cross-validation on the train data to set the value of C. Performance is measured by mean Average Precision (mAP). For the few-shot and zero-shot experiments, the mAP is averaged only over test labels. For the few-shot experiments, we fix $\beta = 1$, see Eq. (11).

We also report the *supervised upper bound* (SUB) performance, obtained by training SVMs on the full ground-truth annotations. The SUB serves as the ideal classifier which could be obtained from this data.

	iCLEF10 [17]	H-SUN [4]	CUB-Att [28]
Nr. train images	8,000	4367	5994
Nr. test images	10,000	4317	5794
Nr. labels	93	107	312
Avg. nr. img / label	925	219	545
Avg. nr. lab / img	12.0	5.34	31.5
SUB in mAP	35.7	28.3	16.6

Table 1. Basic statistics of the three data sets used in our experiments, together with the mAP performance of supervised SVMs.

Data sets. Computing co-occurrences naturally implies that visual concepts must appear frequently together in the image corpus. Although examining a picture will most definitely reveal hundreds of categories present in the frame [29], most often only a handful of those are eventually considered [5]. Since for scarcely labeled datasets co-occurrences will reveal only very little information, we focus on *richly annotated multi-labeled datasets*. We evaluate our methodology on three recent, publicly available, multi-labeled datasets, demonstrating the potential for zero-shot and few-shot recognition.

The *iCLEF10* data set was used in the ImageCLEF 2010 Photo Annotation task[17]. The images are labeled with 93 diverse concepts, containing objects, abstract concepts and aesthetic quality. This allows for computing co-occurrence statistics sufficiently well. On this dataset we obtain 35.7% mAP with our 3K dimensional features. This is somewhat below the 39.0% mAP reported by the challenge winners [17], however their FV used a larger codebook, both SIFT and color features, and spatial pyramids.

The *H-SUN* data set was introduced by [4] for object detection and recognition using hierarchical contextual models. The dataset contains 107 different concepts, most of which are objects. For our experiments we only use the image labelling (not the bounding-box annotations), and we obtain 28.3% mAP. Unfortunately [4] does not provide mAP results, however our SVM results are comparable to the much larger FVs of [15] where 29.8% mAP is reported.

The *CUB-Att* data set refers to the attribute data of the Caltech-UCSD Birds 2011 dataset [28], which contains 200 types of birds and per image attribute annotations. While this dataset is often used for attribute-based prediction, we focus on the multi-label performance using the 312 binary attributes. To the best of our knowledge, we are the first to report attribute performance in mAP, averaged over all labels we obtain 16.9% mAP. To compare our features we rely on the AUC performance over the attributes, in [1] 61.8% AUC is reported using 64K dimensional FV, our 3K dimensional features obtain 59.4% AUC.

	iCLEF10	H-SUN	CUB-Att
Co-Occurrences			
Normalized	24.3	15.0	14.4
Binarized	22.2	15.0	13.3
Square-root	22.2	15.2	13.1
Dice	**25.7**	**18.5**	**14.7**
Positive & Negative Co-occurrences			
Normalized	27.1	14.9	**16.7**
Binarized	22.6	12.7	13.5
Square-root	27.5	15.6	16.5
Dice	**28.4**	**17.3**	16.4
Regression on co-occurrences			
Co-Oc Normalized	28.0	19.1	16.2
Co-Oc Dice	27.5	18.6	16.2
P&N Normalized	**30.7**	20.9	**16.7**
P&N Dice	30.4	**21.1**	**16.7**

Table 2. Overview of zero-shot recognition in a leave-one-out setting, evaluated on the three data sets. We evaluate several similarity measures based on the co-occurrence statistics.

4.2. Zero-shot learning

Co-occurrence similarities. In our first experiment we evaluate the performance of the different similarity metrics. We consider the performance of the similarities (i) when using positive co-occurrences alone, and (ii) when using positive and negative co-occurrences. For both settings we also evaluate the performance when employing regression.

In this experiment, each label is used to estimate a zero-shot classifier in a leave-one-out manner, *i.e.*, when using all $m-1$ other labels. We evaluate the performance by averaging AP over all labels. The results for the three data sets are shown in Table 2.

From the results we observe that our most simple models, where zero-shot classifiers are estimated just on the positive co-occurrences of labels already obtains reasonably good performance. From the different similarities, the Dice coefficient seems to perform best, and clearly outperforms the binarized co-occurrences which are most similar to attributes. For the iCLEF10 and H-SUN data sets the performance compared to SUB, decreases by about 10% mAP, while for the CUB-Att the decrease is less than 2% mAP.

Furthermore, we observe that we could improve these results considerably by adding more advanced co-occurrence statistics. By including both positive and negative co-occurrences we increase performance up to 3% mAP on iCLEF10. Finally, using regression to estimate the zero-shot classifiers yields the best performance on all datasets. On CUB-Att , the P&N Regression models (16.7% mAP) are even slightly outperforming SUB (16.6%).

For the remaining experiments we will report performance obtained by the Dice coefficient, either when us-

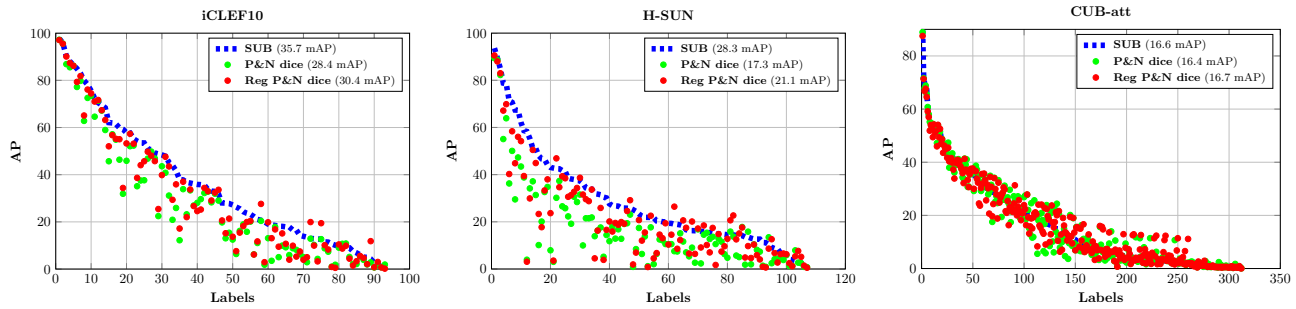

Figure 2. Illustrating the per label performance measure in average precision (AP), comparing the P&N Dice model and, the Regr. P&N Dice model. Note how the zero-shot classifiers approach the SUB performance, and occasionally even outperforms this upper bound.

ing co-occurrences (Co-Oc Dice), positive and negative co-occurrences (P&N Dice), or regression on P&N (Reg. P&N Dice), and we also report results from regression on the P&N normalized similarity (Reg. P&N Norm).

To better understand the performance of our zero-shot model, we visualize the average precision (AP) per label in Figure 2. We compare the supervised upper bound to the co-occurrence from P&N Dice and its regressed variant (Reg. P&N Dice). From the results we observe that, surprisingly, on each dataset there are concepts which could be better described by our zero-shot model than by the supervised upper bound. For example, on the H-SUN dataset, the labels that improve more than 5% over the SVMs are *showcase*, *rug*, *vase*, *oven*, *armchair*, and *poster*. We hypothesize that these labels are less supported by visual examples in the data set, but do co-occur strongly with related labels.

Disjoint label sets. In the second experiment we evaluate the performance when the set of known labels is completely disjoint from the labels used for evaluation. We consider two scenarios, when the set of known labels consist of 75% of the available train labels, and the scenario when it consist of 50% of the labels. In both scenarios we evaluate the performance on a held-out set of 25% of the labels.

We compare the results to the SUB, using all train data, and also to the leave-one-out setting from above, both evaluated just on the held-out labels. Moreover, we adjusted the attribute-based classification approach [12] for multi-label classification: Images are ranked according to the probability for unknown label l. Using $p(l|\boldsymbol{x}) = \frac{1}{p(\boldsymbol{a}^l)} \prod_{k=1}^{K} p(a_k^l|\boldsymbol{x})$, where $\boldsymbol{a}^l$ is the attribute representation of l (defined as the binarized co-occurrence vector), $p(\boldsymbol{a}) = \prod_k p(a_k)$, using the empirical means of training labels and $p(a_k|\boldsymbol{x}) = \sigma(w_k^\top \boldsymbol{x})$, see [12] for details. We denote this as *attributes* in the table.

The results for all three data sets are presented in Table 3. Note that evaluation is performed on the held-out set, causing different SUB mAP values compared to Table 1.

From the results we observe that the performance of our co-occurrences based methods are robust against smaller training sets. For example, the decrease in performance from the leave-one-out setting to using just 50% of the available train data on the iCLEF10 dataset is just 3% mAP. On the CUB-Att the performance remains the same, which is probably due to the high correlations between the labels.

Furthermore, compared to the attribute model of [12] we observe that co-occurrences are notably more accurate, improving up to 7% when using co-occurrences with regression. We explain it by the powerful co-occurrence encoding of visual concepts, compared to binary relevance.

4.3. Web statistics

In this section we consider using the Yahoo search engine and Flickr website to estimate the co-occurrence statistics. Because CUB-Att will not yield interesting results due to the nature of its labels (*e.g.*, *yellow beak* and *white belly*), we perform this experiment on iCLEF10 and H-SUN. For each label and for each pair of labels, we query web, image and Flickr tag search engines to obtain the hit counts.

We have used the regression on the positive and negative Dice coefficients (denoted as Regr. P&N Dice). In Table 4 we show the zero-shot classification results using the hit count based co-occurrence similarities. We observe that the performance of zero-shot recognition is heavily influenced by the estimation of the co-occurrence statistics. In general, using web-statistics decrease performance notably, this result is inline with [21], where similar results are obtained for attribute-based classification.

The web search engine does not result in usable co-occurrence statistics for the natural images in our datasets. The image search engine performs already better, but is still worse than the ones obtained from the data set ground truth annotation. The Flickr tags provide usable co-occurrence statistics, without explicit manual labeling. On iCLEF'10 the Flickr Tag co-occurrences are 4% better than Yahoo image search, and just 5% below the trainset co-occurrences in the L1O setting. For H-SUN, similar observations hold, Flickr Tags are 3% better than Yahoo image search, and just 2% below ground-truth co-occurrences.

We conclude that Flickr tags are a reliable source of co-

Setting	iCLEF10				H-SUN				CUB-Att			
	SUB	L1O	ZS75	ZS50	SUB	L1O	ZS75	ZS50	SUB	L1O	ZS75	ZS50
Nr Train Labels	93	92	70	47	107	106	81	54	312	311	234	156
Baselines												
Supervised Upper Bound	**44.6**	-	-	-	**21.5**	-	-	-	**15.4**	-	-	-
Attributes, following [12]	-	34.3	35.1	33.2	-	12.8	13.0	12.3	-	12.6	12.4	12.3
COSTA												
Co-Oc Dice	-	36.1	35.2	35.3	-	14.5	14.5	12.9	-	13.4	12.6	12.7
P&N Dice	-	38.8	38.6	**37.7**	-	13.7	13.8	10.8	-	14.6	**14.4**	14.0
Regr. P&N Norm.	-	**41.6**	**39.9**	36.7	-	16.7	**16.4**	14.5	-	**15.3**	13.9	14.6
Regr. P&N Dice	-	41.0	39.3	36.7	-	**17.0**	**16.4**	**15.0**	-	15.1	13.7	**14.8**

Table 3. Overview of zero-shot classification performance using co-occurrence statistics, all methods are evaluated on a subset of 25% of the labels. We use various settings, supervised SVM upper bound (SUB), leave-one-out (L1O), and two zero-shot prediction models with disjoint train and test labels (ZS75 and ZS50, using 75% and 50% of the available labels respectively).

	Setting	SUB	L1O	ZS75	ZS50
iCLEF10	**Label Annotations**				
	SUB	**44.6**	-	-	-
	Label Co-oc	-	**41.0**	**39.3**	**36.7**
	Internet search				
	Web hit counts	-	29.0	20.2	24.2
	Image hit counts	-	33.0	24.9	27.4
	Flickr hit counts	-	**36.8**	**30.7**	**31.4**
H-SUN	**Label Annotations**				
	SUB	**21.5**	-	-	-
	Label Co-oc	-	**17.0**	**16.4**	**15.0**
	Internet search				
	Web hit counts	-	9.9	9.8	9.8
	Image hit counts	-	12.7	9.1	9.3
	Flickr hit counts	-	**15.1**	**13.4**	**10.1**

Table 4. Performance when estimating the co-occurrences from web search engines on the iCLEF10 and H-SUN dataset. We use the zero-shot Regr. P&N Dice model.

occurrences *in the wild*, confirming we exploit natural co-occurrences of visual concepts.

4.4. Few-shot learning

In this final experiment we illustrate that the zero-shot model can serve as prior in a few-shot classification setting. This could be beneficial particularly when there are just a few (*e.g.*, up to 8) positive instances per label. Using the setting where we use 75% of the labels as known data, we add a few positive instances for the remaining 25% of the labels. The performance is evaluated over the test set of the latter labels. We run this experiment on all three data sets, using the Regr. P&N Dice model on the ground-truth co-occurrences. For the iCLEF10 and H-SUN dataset we also consider the Flick hit count co-occurrence model als prior.

In Figure 3 we show the results of the few-shot baseline classifier $\boldsymbol{w}_g$ and the results when including the prior $\hat{\boldsymbol{w}}_l = \boldsymbol{w}_g + \beta \boldsymbol{w}_z$, using either the ground-truth or web co-occurrence statistics The value for C is set using 2-fold cross-validation for the few-shot baseline, the extended model uses the same value and we fix $\beta = 1$. This might not be the most optimal setting, especially when training data is plentiful, which could be observed in the results.

From the results we observe that including the prior increases the performance significantly in the few-label range.

5. Conclusion

In this paper we have introduced COSTA for using co-occurrence statistics for zero-shot multi-label image classification. To the best of our knowledge, we are the first to present a model for *multi-label* zero-shot classification. We believe that co-occurrence statistics are a natural way to describe new labels in many real-life scenarios. They describe a new label within a context of related visual concepts.

On three data sets we have shown that co-occurrence statics create powerful zero-shot recognition models. Moreover we have shown that co-occurrences can be obtained from external sources, such as web search engines, confirming that we exploit natural co-occurrences of visual concepts and not just the dataset bias. Finally we have illustrated that the zero-shot model can act as prior in a few-shot classification setting.

We consider the findings in this paper as a starting point for future research. We highlight two possible directions, first the co-occurrence similarities could be defined using higher order semantic relations, *e.g.*, by hierarchical models or Wordnet relations. Second, we could combine different co-occurrence similarities to obtain better zero-shot prediction models, and better priors for few-shot classification.

We conclude that co-occurrence statistics suffice for zero-shot classification.

Acknowledgments. This research is supported by the STW STORY project and the Dutch national program COMMIT.

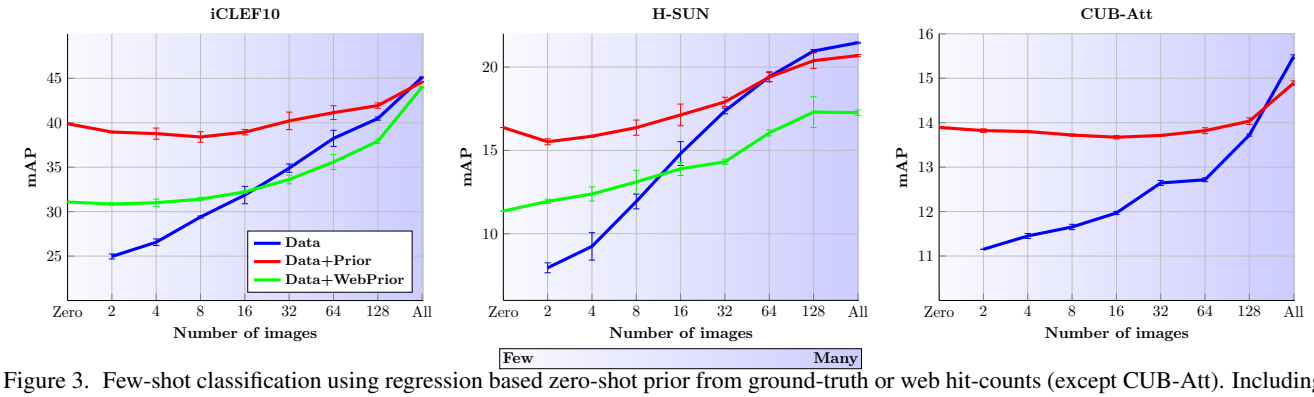

Figure 3. Few-shot classification using regression based zero-shot prior from ground-truth or web hit-counts (except CUB-Att). Including the prior significantly benefits performance, especially in the presence of very few positive images. Note the log-scale on the x-axis.

References

[1] Z. Akata, F. Perronnin, Z. Harchaoui, and C. Schmid. Label-embedding for attribute-based classification. In *CVPR*, 2013. 2, 5

[2] E. Bart and S. Ullman. Cross-generalization: Learning novel classes from a single example by feature replacement. In *CVPR*, 2005. 2

[3] G. Chen, Y. Ding, J. Xiao, and T. Han. Detection evolution with multi-order contextual co-occurrence. In *CVPR*, 2013. 3

[4] M. Choi, J. Lim, A. Torralba, and A. Willsky. Exploiting hierarchical context on a large database of object categories. In *CVPR*, 2010. 3, 5

[5] M. Everingham, L. Van Gool, C. Williams, J. Winn, and A. Zisserman. The Pascal visual object classes (VOC) challenge. *IJCV*, 2010. 2, 5

[6] R.-E. Fan, K.-W. Chang, C.-J. Hsieh, X.-R. Wang, and C.-J. Lin. Liblinear: A library for large linear classification. *JMLR*, 2008. 4

[7] H. Grabner, J. Gall, and L. V. Gool. What makes a chair a chair? In *CVPR*, 2011. 1

[8] B. Hariharan, S. Vishwanathan, and M. Varma. Efficient max-margin multi-label classification with applications to zero-shot learning. *MLJ*, 2012. 3

[9] H. Jégou and O. Chum. Negative evidences and cooccurrences in image retrieval: the benefit of PCA and whitening. In *ECCV*, 2012. 3

[10] H. Jégou, M. Douze, and C. Schmid. On the burstiness of visual elements. In *CVPR*, 2009. 3

[11] L. Ladický, C. R. P. Kohli, and P. Torr. Graph cut based inference with co-occurrence statistics. In *ECCV*, 2010. 3

[12] C. Lampert, H. Nickisch, and S. Harmeling. Attribute-based classification for zero-shot learning of object categories. *IEEE Trans. PAMI*, 2013. 1, 2, 3, 6, 7

[13] F.-F. Li, R. Fergus, and P. Perona. One-shot learning of object categories. *IEEE Trans. PAMI*, 2006. 2

[14] T. Malisiewicz and A. Efros. Beyond categories: the visual memex model for reasoning about object relationships. In *NIPS*, 2009. 1, 3

[15] T. Mensink, J. Verbeek, and G. Csurka. Tree-structured CRF models for interactive image labeling. *IEEE Trans. PAMI*, 2012. 3, 5

[16] T. Mensink, J. Verbeek, F. Perronnin, and G. Csurka. Distance-based image classification: Generalizing to new classes at near-zero cost. *IEEE Trans. PAMI*, 2013. 1, 2

[17] S. Nowak and M. Huiskes. New strategies for image annotation: Overview of the photo annotation task at ImageCLEF 2010. In *Working Notes of CLEF*, 2010. 5

[18] F. Orabona, C. Castellini, B. Caputo, A. E. Fiorilla, and G. Sandini. Model adaptation with least-squares svm for adaptive hand prosthetics. In *IEEE ICRA*, 2009. 4

[19] M. Rastegari, A. Farhadi, and D. Forsyth. Attribute discovery via predictable discriminative binary codes. In *ECCV*, 2012. 2

[20] M. Rohrbach, M. Stark, and B. Schiele. Evaluating knowledge transfer and zero-shot learning in a large-scale setting. In *CVPR*, 2011. 1, 2

[21] M. Rohrbach, M. Stark, G. Szarvas, I. Gurevych, and B. Schiele. What helps where – and why? semantic relatedness for knowledge transfer. In *CVPR*, 2010. 2, 4, 6

[22] M. Sadeghi and A. Farhadi. Recognition using visual phrases. In *CVPR*, 2011. 3

[23] R. Salakhutdinov, A. Torralba, and J. Tenenbaum. Learning to share visual appearance for multiclass object detection. In *CVPR*, 2011. 2

[24] J. Sánchez, F. Perronnin, T. Mensink, and J. Verbeek. Image classification with the fisher vector: Theory and practice. *IJCV*, 2013. 3, 4

[25] V. Sharmanska, N. Quadrianto, and C. Lampert. Augmented attribute representations. In *ECCV*, 2012. 2

[26] 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. 2

[27] T. Tommasi, F. Orabona, and B. Caputo. Safety in numbers: Learning categories from few examples with multi model knowledge transfer. In *CVPR*, 2010. 4

[28] C. Wah, S. Branson, P. Welinder, P. Perona, and S. Belongie. The caltech-ucsd birds-200-2011 dataset. Technical report, Computation & Neural Systems, 2011. 1, 2, 3, 5

[29] J. Xiao, J. Hays, K. A. Ehinger, A. Oliva, and A. Torralba. SUN database: Large-scale scene recognition from abbey to zoo. In *CVPR*, 2010. 2, 5