# Transfer Learning via Attributes for Improved On-the-fly Classification

Praveen Kulkarni, Gaurav Sharma, Joaquin Zepeda, Louis Chevallier
Technicolor
http://research.technicolor.com

## Abstract

*Retrieving images for an arbitrary user query, provided in textual form, is a challenging problem. A recently proposed method addresses this by constructing a visual classifier with images returned by an internet image search engine, based on the user query, as positive images while using a fixed pool of negative images. However, in practice, not all the images obtained from internet image search are always pertinent to the query; some might contain abstract or artistic representation of the content and some might have artifacts. Such images degrade the performance of on-the-fly constructed classifier.*

*We propose a method for improving the performance of on-the-fly classifiers by using transfer learning via attributes. We first map the textual query to a set of known attributes and then use those attributes to prune the set of images downloaded from the internet. This pruning step can be seen as zero-shot learning of the visual classifier for the textual user query, which transfers knowledge from the attribute domain to the query domain. We also use the attributes along with the on-the-fly classifier to score the database images and obtain a hybrid ranking. We show interesting qualitative results and demonstrate by experiments with standard datasets that the proposed method improves upon the baseline on-the-fly classification system.*

## 1. Introduction

Image classification is one of the central problems of computer vision. Many works *e.g.* [11, 6] have been proposed to address the problem of classification of scenes and objects. The problem has been traditionally addressed in a supervised learning scenario where the task is to learn an image classifier when some *positive* examples i.e. images containing the scene or object of interest are given. However, systems developed with such assumptions suffer from obvious limitations: (i) the number of scene or object categories is very large and (ii) annotating, let alone conceiving all textual queries that users might be interested in, is impractical. To address these limitations Chatfield and Zisserman [1] proposed to learn visual classifiers for the user provided textual queries, *on-the-fly*. In their proposed method, they first used the user query to search for images on the internet using an image search engine. Then they used the top images returned by the search as the visual examples of the query and trained the corresponding visual classifier against a fixed set of generic negative images. They then finally used this classifier to obtain a ranking of the images in the database.

Figure 1. Some top ranked images retrieved by Google image search for query 'dog'.

However, relying solely on internet image search for on-the-fly classification leads to the following problem in practice. As shown in Fig. 1, quite a few of the top ranked images returned for even simple queries contain (i) objects with artifacts, or (ii) objects in rare and/or unusual poses/appearances/viewpoints or (iii) artistic/'professional' images with misleading (*e.g.* white) background context. These images degrade the performance of the on-the-fly classifier. For example, Fig. 2 shows the retrieval results for three different animal queries, with the classifier trained with one such image as the only positive example ( [14]). We can see that clearly such images are not suitable for the task. Thus one of the basic task addressed by the present work is automatic filtering of such images towards the goal of improving on-the-fly classification.

We propose to do such filtering by performing *transfer learning*. We use the domain of *attributes* e.g. 'furry', 'has four legs', 'spotted' etc. and transfer knowledge from here to the domain of on-the-fly query based classification. Like many previous works [5, 10, 21] we argue that attributes are useful for visually characterizing a class. A large set

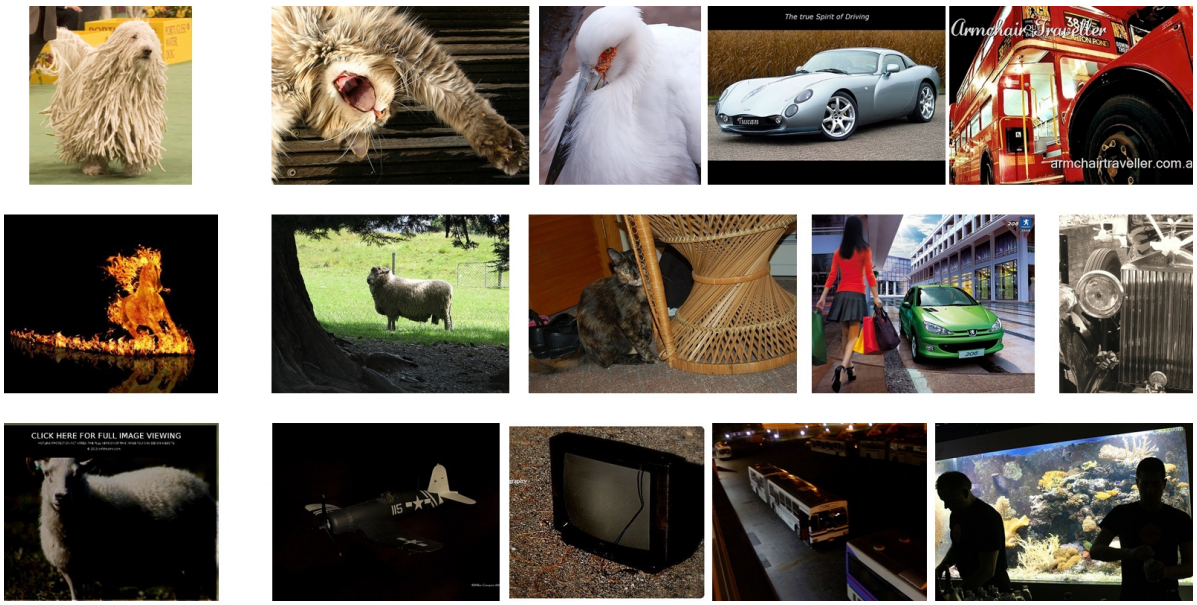


Figure 2. Top retrievals for some of the non-pertinent images returned by Google image search: for each row, the images from Google (on the left) were used to train exemplar SVMs [14] which were in turn used to retrieve images from the Pascal VOC 2007 dataset with some top false positives shown on right. Note how rare appearances, abstract/artistic representations and misleading background leads to poor retrieval *wrt* the queries.

of classes can be potentially covered by combining smaller number of attributes. Hence, if we have annotations for a relatively small set of attributes we can transfer this knowledge to the much larger set of all queries resulting from the combination of these attributes. We note here that although a large number of works using attributes have been reported in the recent past, such use of attributes in the context of on-the-fly classification has never been explored before.

The two main contributions of this paper are for improving on-the-fly classification with the help of attributes. First, we use the attributes to do a *zero-shot* classification on the set of images returned by the internet image search. We discard the images which score low with such attribute based classifier as they are likely to be visually less informative, if not completely wrong or misleading. Second, we use this zero-shot classifier to also score the database images. We combine this score with that obtained by the on-the-fly classifier to obtain a hybrid ranking of the images. We describe our two contributions in more detail in Sec. 3. We show qualitative and quantitative experimental results (Sec. 4) to demonstrate that the proposed approaches improve the baseline on-the-fly system. We now discuss some closely related works in the following section.

## 2. Related Works

Our work is primarily related to image classification, on-the-fly classification, attributes and transfer learning. We now discuss closely related works in the following.

**Image classification** is an important computer vision problem and there are many works on this topic. Many of the current classification systems e.g. [1, 6, 11, 18] are based on the so called bag-of-features representation [2, 19] where local appearance features (e.g. SIFT [13]) are extracted for patches on a dense grid over the image and then vector quantized w.r.t. a codebook. The codebook itself is learned offline by performing some standard clustering algorithm e.g. $k$-means on randomly samples features from the training images. To encode some spatial information Lazebnik et al. [11] proposed to pool over spatial cells i.e. make a spatial pyramid gaining substantial performance. We follow these works here and use this image representation.

**On-the-fly classification** is an extension of supervised image classification to the case of arbitrary user-specified queries [1]. It addresses the limitation of standard classification, i.e. necessity of annotated positive images pertinent to the query, by constructing a positive set of images on-the-fly by querying image search engines on the internet. After such positive images are obtained standard classification algorithms are applied.

**Attributes** have become quite popular in computer vision. They have been used to describe objects [5], and improve image classification performance [21]. An interesting use of attributes was shown in zero-shot learning [10, 28] where classifiers were learnt without any images for the class via attributes *e.g.* Yu and Aloimonos [28] learned

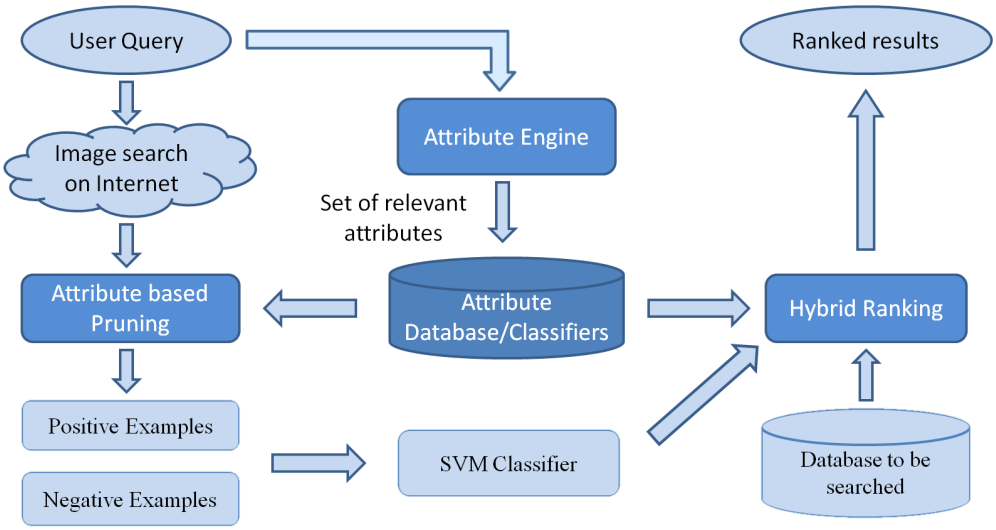


Figure 3. Block diagram of the proposed system. Our contributions are highlighted (see Sec. 3).

the generative attribute models which were used as priors. These attribute priors were shown to improve image classification performance in zero-shot and/or one-shot learning framework.

A lot of work has also been done to represent the images in low dimensional attribute space [12] with the coefficients of the feature vector as the scores of the attribute classifiers. Wang et al. [25] proposed to learn image similarity from Flickr image groups. Vogel and Schiele [24] represented images by concatenating the local semantic descriptions into one large global descriptor and then used them for retrieval of natural scenes. Torresani et al. [22] used large number of weakly trained attribute classifier outputs to represent an image and used it for classification. Kumar et al. [9] used attribute classifiers (e.g. gender, race, hair color) for the task of face verification i.e. to tell if two given faces are of the same person or not.

In addition to image representation using attributes, a lot of work in cognitive science [7, 15, 16, 20] has focused on understanding how humans perceive the relations between attributes and objects.

**Transfer learning** is defined as using the knowledge learned in one task to new task which share some statistical relationship. Duan et al. [3] learned a target classifiers using a set of independent auxiliary classifiers learnt in some other domain. Yang et al. [27] proposed to address two problems in classifier adaptation. First, adapting the multiple classifiers learnt in auxiliary domain to do classification in target domain and second, learning the selection criteria for best classifier in auxiliary domain. Duan et al. [4] proposed to learn both cross domain kernel function and also robust SVM classifier in video concept detection. Wu and Dietterich [26] worked in SVM framework and used the kernel derived from auxiliary domains, containing large amount of training data, in the target domain containing very less training data.

Transfer learning has also been applied to attribute based image classification. Russakovsky and Fei-Fei [17] proposed to obtain the visual connection between object categories based on transfer learning. They started with learning 20 visual attributes from ImageNet data and used these attributes to find the connection between the object categories.

## 3. Approach

In the following, first we set the background context by describing the on-the-fly classification [1] method briefly. We then describe our proposed attribute-based positive image set pruning via zero-shot. Finally, we describe our proposed hybrid ranking system obtained by combining the on-the-fly classifier and an attribute-based zero-shot classifier. Fig. 3 gives the overall block diagram of the proposed system with blocks corresponding to our novel contributions highlighted.

### 3.1. On-the-fly classification

On-the-fly classification is a method of image retrieval based on arbitrary user queries. It uses standard binary supervised classification setup with the positive images obtained from internet image search while keeping the negative images fixed (a set of generic negative images). As the user provides a query, the system makes the same query to an image search engine on the internet. Then the system downloads the top images returned for the query as the positive examples of the query. These are used to learn a SVM classifier which is in turn used to score the database im-

ages. The number of positive images obtained is small and the features for the negative set is already cached, hence the overall features are obtained in reasonable time. A linear SVM classifier is usually learned using stochastic gradient descent which is also fast. Finally, the pertinence score, for the database images, is just a dot product between the learnt classifier and the previously computed and cached features of the database images.

### 3.2. Attribute-based pruning

However, as discussed in the introduction (Sec. 1) the problem with using internet based image search is the risk of obtaining uninformative and/or rare and misleading images as positive examples (Fig. 1). To prune out such images we propose to use transfer learning based on an auxiliary domain of attributes. We propose to do this by constructing a *zero-shot* classifier, inspired by the work of Lampert et al. [10], by mapping the query to a subset of attributes in the attribute dataset. Such mapping could be obtained by a textual analysis system. Using the attributes we learn a zero-shot classifier as follows. We learn a set of attribute classifiers $\{a_i | i \in \mathcal{A}\}$, where $\mathcal{A}$ is the set of attributes, offline. Given a test query $q$, we obtain a set of attributes $\mathcal{A}_q$ corresponding to the query. We then calculate the score matrix for the attribute classifier for all the images $X^+$ downloaded from the internet:

$$A = (a_i^{\mathrm{T}} x_j)_{ij} \forall i \in \mathcal{A}_q, x_j \in X^+.$$

In order to bring the scores of the different attribute classifiers into the same scale, we then normalize the attribute score matrix along its rows. Letting $\mu_i$ and $\sigma_i$ denote, respectively, the mean and variance of the entries in the $i$-th row of $A$, the normalized score matrix $A'$ is given by

$$A' = \left(\frac{A_{ij} - \mu_i}{\sigma_i}\right)_{ij} \tag{1}$$

Finally the attribute based score is given by the sum over all the attributes for positive images. This is our zero-shot attribute based classifier score as it was derived by transferring knowledge from the auxiliary attribute domain, and without the need for training images for the query. The resulting scores are indicative of the presence of the attributes related to the query in the corresponding downloaded images. We hence discard the lowest scoring $k$ images, as they are likely uninformative. Using these pruned images as positive examples, we learn a linear SVM classifier $w$. We then compute the pertinence of the images in the database $X^R$, w.r.t. on-the-fly classifier, as

$$s^o = w^{\mathrm{T}} X^R. \tag{2}$$

### 3.3. Hybrid ranking with attributes and on-the-fly classifier

The attribute based zero-shot query classifier can also be used to test the pertinence of the images in the retrieval database $X^R$. To this end, we build a score matrix $B = (a_i^{\mathrm{T}} x_j)_{ij} \forall i \in \mathcal{A}_q, x_j \in X^R$ using the same attributes $\mathcal{A}_q$ used for the positive set pruning process. The score matrix $B$ is also centered and normalized row-wise as in (1) to produce $B'$. The summation over the columns of $B'$ (*i.e.* scores from only relevant attributes $\mathcal{A}_q$) given by

$$s^a = \mathbf{1}^{\mathrm{T}} B' \tag{3}$$

is score of the database images w.r.t. the query, based on zero-shot attribute classifier.

We propose a hybrid ranking score that combines the two pertinence scores $s^o$ and $s^a$. To do so, we need to bring the two scores into the same scale. Letting $(\mu_1, \sigma_1)$ and $(\mu_2, \sigma_2)$ denote, respectively, the mean and variance of $s^o$ and $s^a$, we define the hybrid score for image $k$ as

$$\alpha \frac{s_k^o - \mu_1}{\sigma_1} + (1 - \alpha) \frac{s_k^a - \mu_2}{\sigma_2} \tag{4}$$

The weight $\alpha$ controls the relative importance of the attribute-based score $s^o$ and the score $s^a$ of on-the-fly classifier with pruned images.

## 4. Experimental results

We validate our method on publicly available Pascal VOC 2007 [6] dataset. We restrict our domain of interest to the animal classes. For the auxiliary domain of attributes for transferring knowledge we use the Animals with Attributes [10] dataset. We use the original on-the-fly system [1] as the baseline method and show by experiments how our methods improves the baseline. We first give details of the datasets and the implementation and then proceed to show our quantitative and qualitative results.

### 4.1. Datasets used

**Pascal VOC 2007** dataset [6] consists of 9163 images with 20 object categories such as car, bicycle, horse, potted plant. The dataset is split into training, validation and test sets. The test set consists of 4,192 images. Ground truth data is available for the complete test dataset. We use all the test images but restricting the performance evaluation to the domain of animals, we report results using five animal classes as queries i.e. horse, cat, dog, cow and sheep. Note that the other classes are present as distractors in the test set. Performance is evaluated by computing the precision at 10 images, that at 50 images and the average precision for each class, as well as the mean of the three metrics for all the five classes.

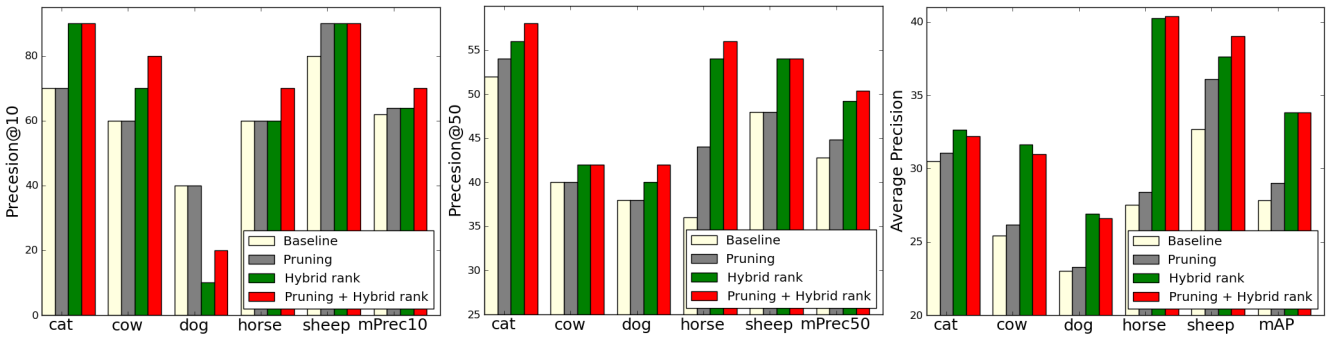


Figure 4. The performances (precision at 10 and 50 and average precision) for all the animal classes, along with mean performances, for the proposed method vs. the baseline of on-the-fly classification [1]. See Sec. 4.3 for discussion.

Figure 5. Top false positive for 'dog' class. We see that the results obtained are dominated by animals with closely related attributes like 'furry', 'long legs'. See Sec. 4.3 for more discussion.

Table 1. The attributes used for the five animal queries.

| Query | Attributes |
|---|---|
| horse | furry, big, toughskin, hooves, longleg, longneck, fast, strong, agility, quadrapedal, vegetation, grazer, plains, fields |
| dog | spots, furry, paws, claws, lean, longleg, fast, strong, quadrapedal, active, inactive, plains, fields, ground, fierce, solitary, newworld |
| sheep | furry, bulbous, hooves, mountains, ground, timid |
| cat | furry, small, quadrapedal, weak, active, inactive, agility, hunter, newworld |
| cow | patches, spots, toughskin, hooves, horns, big, quadrapedal, vegetation, grazer, plains, fields, ground, group |

**Animals with Attributes** dataset [10] consists of 30475 images. There are in total 50 animal classes with at least 95 images of each class. Annotations are also provided for 85 attributes related to the animals. We use this database as the source of auxiliary knowledge to be transferred to the query classification domain. Kemp et al. [8] computed a matrix with values specifying the relative strength of association of attributes with object categories. This matrix was built based on feedback by human subjects on association strength between 50 animal classes with 85 attribute categories. We use this matrix in our case to train our attribute classifiers and for the attributes to be used for each animal query, Tab. 1 gives the list of attributes used for each of the five animal queries.

### 4.2. Implementation details

**Internet image search.** We use Google Image search via the publicly available API to obtain image results obtained for a given textual query. Once the results are obtained we download the images in the search results with a timeout threshold. On an average we download about 85 images per query due the limitation imposed by the API and the timeout threshold.

**Bag-of-features.** We represent images similar to [1] with bag-of-features histograms. We use densely sampled gray scale SIFT features extracted at 4 scales with step size of 3 pixels. We use VLfeat library [23] for extracting SIFT features. We learn a visual codebook of size 4,000 using randomly sampled SIFT features from the Pascal VOC 2007 $train + val$ dataset. We use nearest neighbor based hard assignment of SIFT features to codebook vectors. Finally, we use three level spatial pyramid [11] by dividing the image into 1x1, 3x1 and 2x2 spatial grid.

**Attribute based transfer.** For the zero-shot classifier based pruning we discard the bottom $k = 8$ images and for the hybrid scoring, the weight parameter is set to $\alpha = 0.3$, both parameters were set based on validation experiments.

### 4.3. Quantitative results

Fig. 4 shows the results of the proposed methods vs. the baseline of on-the-fly classification scheme of Chatfield and Zisserman [1]. We can make the following observations. First, the attribute based pruning of test images improves

the performance over the baseline by a modest amount specially at the higher end of recall (precision at 50 and mAP). This is consistent for all the classes. Second, doing hybrid attribute and on-the-fly classification based ranking gives large performance improvements again at the higher end of recall. Third, doing both pruning and then hybrid ranking improves performance at lower end of the recall.

The proposed method improves the baseline in all cases except the 'dog' query, where the number of true positives among the top 10 retrieved images decreases from 4 to 2. We analyzed the results in this case and found that the top retrieved images for this case were those of animals with very closely related attributes e.g. 'furry' and 'has four legs'. Fig. 5 shows some of the top false positives. However, we note that the performance recovers at higher recall e.g. the precision at 50 and the average precisions improve for dog class as well.

**Processing times.** The time (on a single core) taken by different steps are as follows. Downloading 85 images (for a query) takes 4s using the Google Search API. Feature extraction takes $\sim 3s$ per image, (ii) SVM learning takes $\sim 6s$ and scoring the database images (which is matrix dot product and sum) is negligible. When multiple cores are used the system is reasonably responsive.

### 4.4. Qualitative results

Fig. 6 shows some qualitative results for our system. Each row corresponds to an animal query, on the left the images retained, for training the on-the-fly system, and on the left the discarded images are shown. We can see that the images which are more natural and are pertinent to the queries are retained by the proposed method. While the image which have either rare object appearance/pose or uninformative/misleading background or are abstract/artistic rendering of the animals have been discarded.

## 5. Discussion and conclusion

In the present paper we presented a method to use attributes to improve the classification performance of on-the-fly [1] classification for retrieval. We showed that transferring knowledge from the attribute domain to the query domain is effective in pruning out images containing (i) objects in rare or unnatural poses, (ii) objects on uninformative/misleading background and/or (iii) artistic or abstract rendering of the objects. We also proposed a hybrid ranking system which, along with the on-the-fly classification, takes the attribute classifiers into account. We showed by experiments on standard publicly available datasets that our methods improves upon the baseline.

The attribute engine, which maps the query to a set of relevant attributes, was assumed to be given in the present paper. The design of an automatic attribute engine is a challenging future work that we would like to pursue. Also, the attribute dataset was assumed to be given, fixed and annotated. It would also be interesting to explore creation or extension of such attribute datasets on-the-fly as well.

## References


[1] K. Chatfield and A. Zisserman. VISOR: Towards on-the-fly large-scale object category retrieval. In *ACCV*, Lecture Notes in Computer Science. Springer, 2013. 1, 2, 3, 4, 5, 6

[2] G. Csurka, C. R. Dance, L. Fan, J. Willamowski, and C. Bray. Visual categorization with bags of keypoints. In *Intl. Workshop on Stat. Learning in Comp. Vision*, 2004. 2

[3] L. Duan, I. W. Tsang, D. Xu, and T.-S. Chua. Domain adaptation from multiple sources via auxiliary classifiers. In *ICML*, 2009. 3

[4] L. Duan, I. W. Tsang, D. Xu, and S. J. Maybank. Domain transfer svm for video concept detection. In *CVPR*, 2009. 3

[5] I. Endres, D. Hoiem, A. Farhadi, and D. Forsyth. Describing objects by their attributes. In *CVPR*, 2009. 1, 2

[6] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The PASCAL Visual Object Classes Challenge 2007 (VOC2007) Results. http://www.pascal-network.org/challenges/VOC/voc2007/workshop/index.html. 1, 2, 4

[7] T. Hansen, M. Olkkonen, S. Walter, and K. R. Gegenfurtner. Memory modulates color appearance. *Nature neuroscience*, 9(11):1367–1368, 2006. 3

[8] C. Kemp, J. B. Tenenbaum, T. L. Griffiths, T. Yamada, and N. Ueda. Learning systems of concepts with an infinite relational model. In *AAAI*, volume 3, page 5, 2006. 5

[9] N. Kumar, A. C. Berg, P. N. Belhumeur, and S. K. Nayar. Attribute and simile classifiers for face verification. In *ICCV*, 2009. 3

[10] C. H. Lampert, H. Nickisch, and S. Harmeling. Learning to detect unseen object classes by between-class attribute transfer. In *CVPR*, 2009. 1, 2, 4, 5

[11] S. Lazebnik, C. Schmid, and J. Ponce. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In *CVPR*, 2006. 1, 2, 5

[12] L.-J. Li, H. Su, L. Fei-Fei, and E. P. Xing. Object bank: A high-level image representation for scene classification & semantic feature sparsification. In *NIPS*, 2010. 3

[13] D. G. Lowe. Object recognition from local scale-invariant features. In *Proceedings of the International Conference on Computer Vision-Volume 2 - Volume 2*, ICCV '99, pages 1150–, Washington, DC, USA, 1999. IEEE Computer Society. 2

[14] T. Malisiewicz, A. Gupta, and A. Efros. Ensemble of Exemplar-SVMs for object detection and beyond. In *ICCV*, 2011. 1, 2

[15] D. Osherson, E. E. Smith, T. S. Myers, E. Shafir, and M. Stob. Extrapolating human probability judgment. *Theory and Decision*, 36(2):103–129, 1994. 3

[16] D. N. Osherson, J. Stern, O. Wilkie, M. Stob, and E. E. Smith. Default probability. *Cognitive Science*, 15(2):251–269, 1991. 3

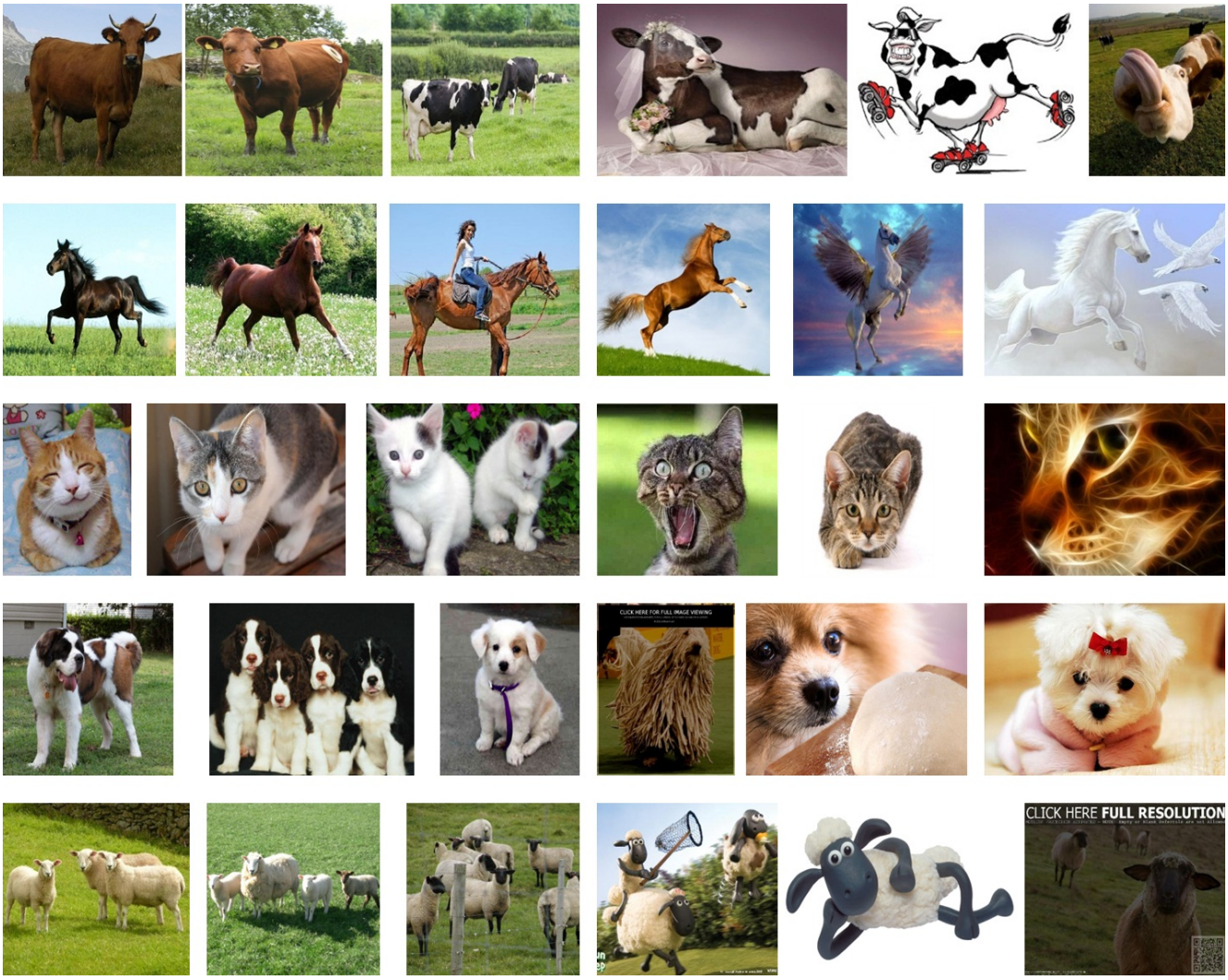


Figure 6. Example of images from Google image search which are (left three) retained by our system as pertinent to the query (animals i.e. cow, horse, cat, dog, sheep, from top to bottom) and (right three) discarded by our system as being uninformative. We see that the images which are more natural and pertinent for the class queries and retained while those which have objects on rare or unnatural poses, or objects on uninformative/misleading background or artistic or abstract rendering of the objects are discarded by the system.

[17] O. Russakovsky and L. Fei-Fei. Attribute learning in large-scale datasets. In *Trends and Topics in Computer Vision*, pages 1–14. Springer, 2012. 3

[18] G. Sharma, F. Jurie, and C. Schmid. Discriminative spatial saliency for image classification. In *CVPR*, 2012. 2

[19] J. Sivic and A. Zisserman. Video Google: A text retrieval approach to object matching in videos. In *ICCV*, 2003. 2

[20] S. A. Sloman. Feature-based induction. *Cognitive psychology*, 25(2):231–280, 1993. 3

[21] Y. Su and F. Jurie. Improving image classification using semantic attributes. *IJCV*, 100(1):59–77, 2012. 1, 2

[22] L. Torresani, M. Szummer, and A. Fitzgibbon. Efficient object category recognition using classemes. In *ECCV*, pages 776–789. Springer, 2010. 3

[23] A. Vedaldi and B. Fulkerson. VLFeat - an open and portable library of computer vision algorithms. In *ACM Multimedia*, 2010. 5

[24] J. Vogel and B. Schiele. Semantic modeling of natural scenes for content-based image retrieval. *International Journal of Computer Vision*, 72(2):133–157, 2007. 3

[25] G. Wang, D. Hoiem, and D. Forsyth. Learning image similarity from flickr groups using stochastic intersection kernel machines. In *ICCV*, pages 428–435. IEEE, 2009. 3

[26] P. Wu and T. G. Dietterich. Improving svm accuracy by training on auxiliary data sources. In *ICML*, 2004. 3

[27] J. Yang, R. Yan, and A. G. Hauptmann. Cross-domain video concept detection using adaptive svms. In *ACM Multimedia*, pages 188–197. ACM, 2007. 3

[28] X. Yu and Y. Aloimonos. Attribute-based transfer learning for object categorization with zero/one training example. In *ECCV*, pages 127–140. Springer, 2010. 2