# Swap Retrieval: Retrieving Images of Cats When the Query Shows a Dog

Amir Ghodrati, Xu Jia, Marco Pedersoli, Tinne Tuytelaars
KU Leuven, ESAT - PSI, iMinds
Leuven, Belgium
firstname.lastname@esat.kuleuven.be

## ABSTRACT

Query-by-example remains popular in image retrieval because it can exploit contextual information encoded in the image, that is difficult to express in a traditional textual query. Textual queries, on the other hand, give more flexibility in that it's easy to reformulate and refine a text query based on initial results.

In this work we make a first step towards getting the best of both worlds: we use an image to specify the context, but let the user specify a related category as main search criterion. For instance, starting from an image of a dog in a certain situation/context, the goal is to find images of cats with a similar situation/context.

We present an evaluation scheme for this new and challenging task, which we call swap retrieval, and use it to compare various methods. Results show that standard query-by-example techniques do not adapt well to the new task. Instead, techniques based on semantic knowledge extracted from textual descriptions available at training time perform reasonably well, although they are still far from the performance needed for practical use.

## Categories and Subject Descriptors

H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval—*Retrieval models*; I.2.6 [**Artificial Intelligence**]: Learning—*Concept learning*; I.5.4 [**Pattern Recognition**]: Application—*Computer Vision*

## Keywords

Image Retrieval, Attribute (Concept) Discovery, Multimodal Content Analysis

## 1. INTRODUCTION

Image-based retrieval (aka query-by-example) is an important and well-studied problem in computer vision [30, 3, 10, 20, 9, 6]. However, it is not that flexible, especially when compared with text-based retrieval. When using text-based queries for retrieval, the search for images of a certain topic of interest can be an interactive process, where the user refines the query based on the results given by the retrieval algorithm. When searching based on images, on the other hand, a query cannot be easily refined, as altering an image is not straightforward. Image editing tools (e.g. pasting a cropped part of another image) are cumbersome and typically generate disturbing artefacts that affect the search results.


Figure 1: Category-swap image retrieval. Given the query image of a dog with a hat and the user input "- dog + cat", the goal of this work is to retrieve images of a cat with a hat.

At the same time, text-based retrieval also has its limitations. First, most text-based retrieval schemes rely on associated text (e.g. in the form of filenames, image captions or descriptions). In case that text is not available for candidates, the textual query needs to be "translated" into visual description– typically done using pre-trained classifiers for various concepts, a process that, in spite of a lot of progress, is still quite noisy. Second, and this is what we will focus on in this paper, there are aspects that cannot be easily represented with a textual query. For instance, the whole content of an image, such as specific color, spatial layout and composition, or the conditions under which the picture was taken cannot be easily specified by text. However, such information can be easily conveyed by an image. In this respect, letting the method figure out what are the important characteristics of an image and retrieve similar ones can give more satisfactory results than starting from a text-based query.

In this paper, we propose a novel retrieval task where the query is a combination of image and text instead of either image or text alone. It explores the possibility of using a more advanced and human-friendly query for image retrieval, which not only brings more flexibility to the user, but also takes advantage of additional information contained in the query image. Given a query image from one category, a set of images from a different category but semantically as similar as possible to the original image are to be retrieved. The new task can require semantic understanding of visual data, for instance knowing what happens in the scene and which objects participate in an action. We define this task as *category-swap image-based retrieval*, or *swap retrieval* in short. In summary, in this paper we propose, discuss and evaluate a task that extends the standard image-based query towards a more semantic and compositional representation.

We adapt and evaluate several possible solutions using existing methods from the literature. Starting from a feature representa-

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.
*ICMR'15,* June 23–26, 2015, Shanghai, China.
Copyright © 2015 ACM 978-1-4503-3274-3/15/06 ...$15.00.
DOI: http://dx.doi.org/10.1145/2671188.2749373.

tion based on DeCAF [5] or Classemes [32], the simplest solution is to directly compare images of two different object categories in the feature space. We also test a domain adaptation based technique [7] where the domains are the two different object categories and a metric learning method [4], where similarity and dissimilarity are computed based on textual descriptions associated to the image. We show that those methods tend to perform poorly and therefore are not suitable for this new task. Thus, we propose an approach based on visual attribute classifiers learned from textual descriptions.

For evaluation we select some pairs of classes from the recently released COCO dataset [19]. In the experiments we mainly focus on the pair dog/cat because it is one of the most interesting and familiar and there are many contexts where both pets can fit but also others that are unique for only one of the two (*e.g.*a dog catching a frisbee). Additionally, we also experiment with cow/sheep and chair/toilet. As the task involves a semantic understanding of the image and may be subjective, for a proper evaluation we resort to two evaluation strategies: the first ranks the images based on text similarity (using the associated textual descriptions provided in the dataset), while the second is based on human evaluation.

The main contributions of this work are: first, we propose swap-retrieval, a new and challenging way of querying data, that provides more flexibility to standard image-based retrieval; Second, two meaningful evaluations of the task are explored, one based on the given image descriptions and the other based on collected human annotations; Third, we explore several possible solutions using existing methods, and further propose a new method which leverages textual descriptions at training time to learn a more semantic representation of the images.

The remainder of the paper is structured as follows: in sec. 2 we present and analyze previous work on similar topics. Sec. 3 introduces swap retrieval in a more formal way, as well as a methodology for its evaluation. Several methods applied to the task are first introduced in sec. 4 and then evaluated in the experimental results in sec. 5. Sec. 6 concludes the paper.

## 2. RELATED WORK

**Image Retrieval.** In a traditional query-by-example image retrieval task, the goal is to retrieve relevant images from a dataset given a query image. Relevance in this context implies images with similar visual content. Many works focused on improving the accuracy and efficiency of large-scale image search either for a specific object instance [23, 6, 9, 24] or at category level [1, 34, 28, 8, 12, 33]. Different from the traditional image retrieval task which aims at finding visually similar images, the new task presented in this paper aims at identifying images with similar content but containing an object from a different category. Concept or attribute based representations seem beneficial for this new task. Leveraging attribute representations for image retrieval and related tasks have been investigated in [15, 31]. Kumar *et al.* [15] use describable visual attributes for face verification and image search. Attributes that correspond to hair, eyes and nose are combined to produce semantic descriptions at multiple levels such as categories and object instances. Siddiquie *et al.* [31] propose an approach for image retrieval with multiple attributes as query. The retrieval performance is improved by modeling the correlation between different attributes. For interactive image search, relative attributes [21] provide a powerful tool. Kovashka *et al.* [14, 13] propose a framework for image search with relative attribute feedback, which guides a user through a coarse-to-fine strategy.. The most related work is probably Bi-Concept Search from Li *et al.* [18], where they represent a query with two tags, one for the object and one for the scene. Instead, in our work we express the object with a tag but represent the scene/context with the query image.

**Automatic Attribute Discovery.** Our work is also related to work on "automatic attribute discovery". Yu *et al.* [36] develop a general framework for attribute-based image retrieval by leveraging a large pool of weak attributes. Weak attributes are defined as a collection of mid-level image representations which are learned in various ways and may not have semantic meanings. Li *et al.* [17] automatically discover and model groups of objects which are composites of objects that have consistent spatial layout and interactions across images. The discovered groups of objects act as distinct visual patterns and help object detection and scene recognition. Yu *et al.* [35] propose an approach to automatically select discriminative "category-level attributes" for visual recognition. These methods automatically discover concept or attribute representations using only image information. In contrast, our attributes are learned leveraging both image and text information, hence having clearer semantic meanings.

Recently, there have been several works which automatically discover attributes or concepts using both image and text information. Quattoni *et al.* [25] propose a framework for classification by training a visual classifier for each word in the corpus of captions associated with images. Instead of directly using the output of the word classifiers as features (as in Classemes [32]), they map the output of the word classifiers into a new space. Berg *et al.* [2] treat strings in descriptions accompanying images as attributes and train a visual classifier for each of them. The discovered attributes are further merged and refined to describe an object's visual appearance. In [16], Layne *et al.* learn semantic attributes by clustering internet photos with tags and use them for re-identification. Similar to such approaches, the proposed method also automatically discovers attributes by mining both image and text data. However, the used attributes are different from the above. We use single (based on one word) and paired (based on the co-occurrence of two words) attributes and find that these two kinds of attributes are complementary and the combination of the two provides an effective tool to describe different aspects of an image. Visual phrases which model the interaction between objects are also explored in [17], [26] and [11]. Our work further supports the effectiveness of such visual phrases models.

## 3. PROBLEM DEFINITION

In the proposed task, given a query image, we are interested in identifying images with similar context but containing an object from another category. The query image is meant to provide the context where an object (e.g. 'dog') of a given class appears, while a textual query provides the target class (e.g., 'cat') that a user is interested in.

Although at first sight swap retrieval looks similar to standard image-based retrieval, and one would hope that methods for standard retrieval also perform well on this more challenging task, we have experimentally noticed that swapping the class of the query puts forward higher requirements to semantic image understanding. CNN-based features such as DeCAF [5] have been shown to perform well in computer vision tasks including image retrieval [29]. However, as shown in Fig. 2, their performance in standard image retrieval is much better than in swap retrieval (for the evaluation we use the NDCG30@k metric defined in section 3.2). Therefore, in this work, we are interested in two problems: i) is there a good way to perform this task? and ii) which is the best metric to evaluate the performance of the different methods?

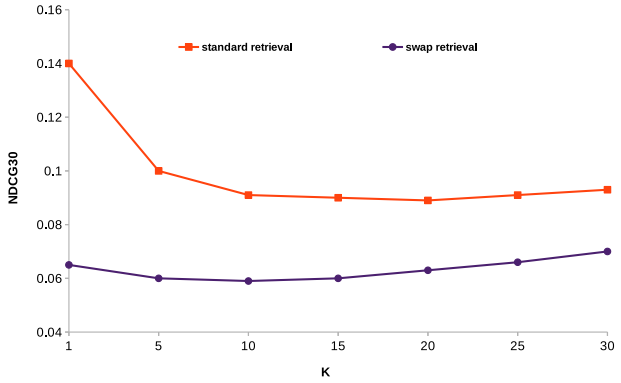


Figure 2: NDCG30@k for standard image retrieval and category-swap retrieval. The swap retrieval is much more difficult than the standard retrieval.

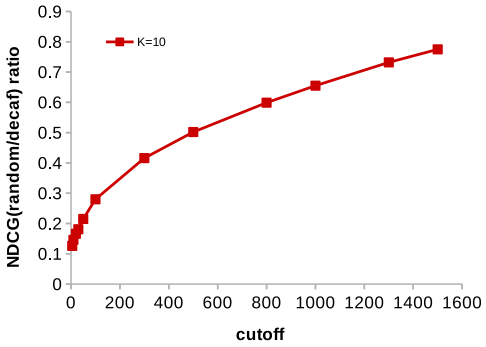


Figure 3: NDCG@k ratio between a random ranking and DeCAF ranking varying the cutoff value. For high values of cutoff, the ratio tends to 1, indicating that the measure loses its meaning.

## 3.1 Ground truth annotations

For evaluation we resort to two complementary strategies. The first is based on human annotations. This is of course the most reliable, but also expensive to obtain. Thus, from the full set of 1400 cats and dogs queries from the COCO dataset, we have selected a random subset of 50 for which we obtained the lists of the swap-retrieval ranking from 10 human annotators. We refer to the supplementary material for more details on the annotation procedure.

The second evaluation strategy relies on the available image descriptions from COCO dataset. This does not require human judgment and therefore can be applied on a large scale, using all 1400 queries. In particular, we rank images based on their textual distance measured based on the bag-of-words representation of descriptions. In the experimental results we compare this evaluation with the annotations provided by humans and show that the two ranks are quite correlated and therefore using the ranking based on textual distance as reference makes sense.

## 3.2 Evaluation metrics

A common metric for measuring ranking performance is the Discounted Cumulative Gain at position k, defined as:

$$\mathrm{DCG_k} = \sum_{i=1}^{k} \frac{2^{rel_i} - 1}{\log_2(i+1)}.$$

DCG and NDCG, its normalized version, measure the ranking quality of a retrieved image based on its graded relevance $rel_i$ and position $k$ in the ranking list. For our specific task, there is only a limited number of relevant images. So if we distribute the relevance among all images, the measurement would be mostly noise because the relevance of similar images is close to that of dissimilar images. An alternative metric is the mean average precision (mAP) but we believe this task is by nature a ranking problem and in AP we cannot give a different relevance to differently ranked images.

Thus, to reduce the aforementioned NDCG problem we propose to assign a graded relevance to the first $m$ similar images only and set the relevance for all others to zero. Figure 3 shows the ratio between the NDCG score of random ranking and the NDCG of the rank obtained using DeCAF features (section 4.1), for varying the cutoff value $m$. When the cutoff increases, the ratio tends to 1 indicating that the measure loses its meaning. Based on this observation, for all experiments we set the cutoff value to 30.

# 4. METHODS

In this section we explore various possible solutions to address the novel task. We first adapt existing well-known techniques for image based classification and retrieval to swap retrieval. Afterwards, we propose an attribute-based method where attributes are automatically mined from images and texts. In all methods, we assume the labels of query and candidate images are given. So for performing swap retrieval, given the query and its label, we retrieve images from the other category.

## 4.1 DeCAF

Our strong baseline is based on DeCAF features [5]. It is an image description obtained from the 6th layer of a convolutional neural network (CNN) trained on ImageNet. In contrast to other features, they indirectly exploit the information of the 1000 ImageNet classes used for training the CNN. Hence they are expected to be more semantic than other low level features and therefore better suited for this task. CNN-based features such as DeCAF have obtained state-of-the-art results on many visual tasks, including image retrieval [29].All the other methods presented here, excluding Classemes, use DeCAF as basic feature representation. For our baseline based on DeCAF, we rank the images according to the similarity score computed as the cosine similarity between the query image of a given category and the candidate images from the swapped category.

## 4.2 Classemes

Classemes [32] are high level features, that are the output of classifiers trained for a large amount of categories. They have been tested in image retrieval with excellent results in [6]. We can consider Classemes as pre-defined attributes and therefore they are considered an interesting baseline for swap retrieval. For ranking the images we use the same procedure as explained in DeCAF, but with the features extracted from Classemes.

## 4.3 Domain Adaptation

Domain adaptation learns a mapping function which aligns the feature space of the source domain to the target domain in order to account for the shift between these two different but related distributions. We can consider swap retrieval as a special case of domain

adaptation, where the domain shift is actually a change in the object category that we are looking for. Pereira and Vasconcelos [22] for instance apply a cross-modal domain adaptation to the task of image retrieval by considering image and text as source and target domains. In our task, we consider the features corresponding to the object category in the query image and the swapped category as source and target domains. Given a query image in the source domain, our goal is finding a similar image in the target domain. We use the method proposed by Fernando *et al.* [7], because of its effectiveness and simplicity. We learn a subspace for each domain and then seek a mapping function that transforms the source subspace into the target subspace. For our task, we learn the subspaces $X_S$ and $X_T$ for each object category using PCA. Hence, the images are ranked based on similarity with the query image in the target subspace, which is defined as: $S(x_q, x_c) = x_q A {x_c}^\top$, where $A = X_S X_S^T X_T {X_T}^\top$.

## 4.4 Metric Learning

A possible way of performing swap retrieval is to learn a metric that properly ranks the images based on the query. This is similar to the work in [27] which employs metric learning for cross-domain adaptation. Here, similar to [27], we use Information Theoretic Metric Learning (ITML) [4] to learn a metric for our task. ITML employs the information theoretic regularization term and learns the metric by minimizing the LogDet divergence. To generate the similarity and dissimilarity constraints suitable to our task, we learn the distance based on the text description associated to images because as already mentioned in section 3.1, text-based distance is semantically more meaningful. For example, for each cat image we choose as similarity constraints 3 images from the dog category whose textual descriptions are the most similar to the cat image, and also the 3 most dissimilar as dissimilarity constraints. These constraints are then used to learn the metric defined by $M$. Given the learned distance metric $M$ we compute the similarity between the query $x_q$ and the candidate image $x_c$ as $x_q M x_c^\top$ and use it for ranking the images for swap retrieval.

## 4.5 Attributes

In this section, our goal is to automatically discover visual attributes from images and associated textual descriptions. Here attributes are defined as semantically meaningful concepts that are able to describe the essence of an image consisting of objects, scenes and actions.

**Attribute selection.** To build an initial candidate list of attributes, we collect a large set of single words from textual descriptions as potential visual attributes. We define the *discriminativity score* of a word $a$ for a pair of classes $c_1$ and $c_2$ as follows:

$$\mathbf{discr}(a) = \left| \frac{1}{|T_{c_1}|} \sum_{t \in T_{c_1}} \mathbf{f}_t(a) - \frac{1}{|T_{c_2}|} \sum_{t \in T_{c_2}} \mathbf{f}_t(a) \right|, \quad (1)$$

where $T_{c_i}$ denotes the set of text descriptions for class $i$, $|T_{c_i}|$ is its cardinality and $\mathbf{f}_t(a)$ is the frequency of attribute $a$ appearing in the text description $t$. Since we use a bag-of-words representation to encode the textual description, intuitively, Eq.1 selects words that occur frequently in one class but not in the other. We select the top scoring K words and add those that occur more than 20 times in the training set to the candidate attribute set.

For this set of candidate attributes, we measure the *visual compactness* of each attribute in terms of variance of the images ($l_2$ norm of the vector composed by the variance of each dimension of the CNN representation) containing that attribute. At this stage, we select the 60% of the attributes with lower variance (visually more compact). From these attributes, we greedily select those that are far from each other in visual space using algorithm 1. That is, if its distance to the closest attribute in the candidate list $A^*$ is higher than $\tau_{min}$ then it is added to candidate list. The distance of two attributes in visual space is defined as:

$$\mathbf{dist}(a_i, a_j) = 1 - \frac{< E(\mathbf{x}_{a_i}), E(\mathbf{x}_{a_j}) >}{|E(\mathbf{x}_{a_i})||E(\mathbf{x}_{a_j})|}, \quad (2)$$

where $x_{a_i}$ denotes the image representation whose textual description contains the attribute $a_i$ and $E(\cdot)$ denotes the mean. The distance here is defined as the negative cosine similarity between images containing the attribute $a_i$ and images contain the attribute $a_j$.

In addition, in algorithm 1, similar to [2], we collect for each attribute a set of visual synonym words or words that co-occur together and we call this list synsets. The attributes are added in the synsets list if their distance is less than fixed threshold $\tau_{max}$. Some examples of learned synsets are shown in table 1 (left column).

**Algorithm 1** Attribute selection

**Input:** Candidate attribute set $A$, whose elements $a_i, i = 1 \cdots N$ are sorted by compactness score, and two thresholds $\tau_{min}$ and $\tau_{max}$
**initialize:** $A^* \leftarrow a_1$
**for** $i = 2 \cdots N$ **do**
  $a_j = \text{argmin}_{a_j \in A^*} \mathbf{dist}(a_i, a_j)$
  **if** $\mathbf{dist}(a_i, a_j) > \tau_{min}$ **then**
    Add $a_i$ to $A^*$
  **else if** $\mathbf{dist}(a_i, a_j) < \tau_{max}$ **then**
    Add $a_i$ to $\mathbf{synset}(a_j)$
  **end if**
**end for**

**Single and paired attributes.** Learning high quality attributes leads to a better representation and therefore better performance. Though single word attributes have good descriptive ability, they cannot capture the relationship between objects. For example one can train "mirror", "car" and "rear" attributes independently using images containing these attributes but they cannot by themselves be representative for "rear mirror of a car". Therefore, similar to "visual phrase" and "bi-concept" in [26] and [11], we also define paired words attributes that aim at representing those more complex concepts that cannot be characterized well by a single word attribute. For simplicity, we abuse the notion and denote the single word attribute as single attribute and paired words attribute as paired attribute. The paired attributes are more suitable for our task since they can capture different relationship between an object and other object/action/scene, hence providing better cues for the swap retrieval.

Learning models for all possible paired-attributes is prohibitive. Also, some pairs are not meaningful and many of them are redundant. Therefore, to extract a set of meaningful paired-attributes we follow the same procedure described above for the single attributes, but this time computing statistics on pairs of words instead of single words. As for single attributes, we also build synsets for paired-attributes as shown in table 1 (right column).

**Attributes for swap retrieval** For swap retrieval we split the paired-attributes into 2 sets: class-specific paired attributes and common paired attributes. In class-specific paired attributes one of the words in the pair corresponds to the class label. For example ⟨cat,bed⟩

| single synset | paired synset |
|---|---|
| luggage, suitcase | ⟨motorcycle,man⟩, ⟨motorcycle,riding⟩ |
| television, tv | ⟨keyboard,sleep⟩, ⟨computing + sleep⟩ |
| bathroom, sink | ⟨hat,cat⟩, ⟨hat,head⟩ |
| naps, asleep | ⟨food,table⟩, ⟨eat,table⟩ |
| boys, young | ⟨people,kite⟩, ⟨people,fly⟩ |
| mothers, baby | ⟨boat,cow⟩, ⟨beach,cow⟩ |
| remote, control | ⟨camera,close⟩, ⟨close,stare⟩ |
| bicycling, biking | ⟨grass,lay⟩, ⟨field,lay⟩ |

Table 1: Some example sysnets detected by algorithm 1

is class-specific if the swap retrieval is effectuated on cat/dog. Exploiting the text information, for each class-specific paired attribute in one class, we build the corresponding paired-attribute in the second class and vice versa. That is, if there is a paired-attribute ⟨cat,bed⟩ in the attribute set, the pair ⟨dog,bed⟩ is also added to the attribute set. With such symmetry, given a query, the retrieval can resort to the class-specific attribute detectors and simply by swapping the detector scores of the attributes, we retrieve images with same attributes but from the opposite class. In contrast, those attributes that do not contain the class names but occur in both classes like ⟨pillow,bed⟩, are called common paired-attributes. The common paired-attributes can help discover a similar context between images and thereby improve the retrieval performance.

**Attribute detector.** We learn a linear visual classifier for each visual attribute using images of the synset associated to the attribute. For learning we use a support vector machine with $C$, the parameter controlling the trade-off between the training error and regularization set to 100 in all experiments. In addition, to cope with unbalanced data during learning, we weigh the classes inversely proportional to the number of samples in the classes. Hence, for each attribute we obtain a vector of parameters $w$ whose dot product with the corresponding image features $x$ gives a score that predicts the presence of the attribute in the image.

**Ranking score.** Now, for the swap retrieval we describe an image as a vector of scores $\psi_a(\mathbf{x})$, where its elements are the scores of all single and paired attributes. Thus, the similarity between two images is defined as:

$$\begin{aligned}&\mathbf{sim_a}(\mathbf{x}_q, \mathbf{x}_c) = \psi_a(\mathbf{x}_q)^T \psi_a(\mathbf{x}_c)\\&\psi_a(\mathbf{x}) = [\psi_a^s(\mathbf{x}), \psi_a^p(\mathbf{x})]\\&\psi_a^s(\mathbf{x}) = [\dots, \mathbf{w}_i^s\mathbf{x}, \dots], \quad \psi_a^p(\mathbf{x}) = [\dots, \mathbf{w}_{ij}^p\mathbf{x}, \dots] \qquad (3)\end{aligned}$$

$\mathbf{w}_i^s$ denotes the parameter of an attribute detector for a single attribute $a_i$ and $\mathbf{w}_{ij}^p$ is the one for a paired attribute $\langle a_i, a_j\rangle$. Reformulating Eq.3, we have

$$\begin{aligned}&\mathbf{sim_a}(\mathbf{x}_q, \mathbf{x}_c) = \sum_{i\in A_s} \mathbf{x}_q^T \mathbf{W}_i^s \mathbf{x}_c + \sum_{i,j\in A_p} \mathbf{x}_q^T \mathbf{W}_{ij}^p \mathbf{x}_c\\&\mathbf{W}_i^s = \mathbf{w}_i^s (\mathbf{w}_i^s)^T, \mathbf{W}_{ij}^p = \mathbf{w}_{ij}^p (\mathbf{w}_{ij}^p)^T \qquad (4)\end{aligned}$$

Thus, $\mathbf{W}$ denotes a matrix that computes similarity of two images in the attribute space. Note that $\mathbf{W}$ is not necessarily a valid metric matrix.

# 5. EXPERIMENTS

In this section we first describe the dataset and the experimental setup that we use. Then we evaluate and compare the methods introduced in section 4 and analyze the results in three different pairs of classes. Finally, we evaluate the paired-attributes in a normal retrieval task.

## 5.1 Experimental Setup

We evaluate our methods on the Microsoft COCO (Common Objects in Context) dataset [19]. COCO contains 82783 training images and 40504 validation images, where a majority of images are non-iconic. Each image, is associated with around 5 sentences that describe the image. In total, 80 common objects are labeled in their natural context. We selected three pairs of categories that are suitable for the specific task of category-swap image retrieval, namely (cat,dog), (cow,sheep) and (toilet,chair). For each pair, images that contain either the first or the second class (but not both) are selected. We use the training images for learning our models and perform leave-one-query-out on the validation set to evaluate the ranking performance. Specifically, at each round one image from COCO validation set is regarded as query and the remaining validation images are used as candidates for retrieval. For training we use both the images and their textual description. At test time, given only the image query we assume that the class label of the query is given so the task is retrieving similar images from the other class.

For visual features we use DeCAF. For textual features, we first remove the stop words from the corpus and after stemming the words, unique words are used for building a *tf-idf* weighted bag-of-words representation. Both visual and textual features are $L2$ normalized in all experiments.

## 5.2 Results and discussion

Here we evaluate the proposed methods on the full dataset, using the text similarity as proxy for the ground truth.

**Cats and Dogs.** We first focus on cast and dogs. In figure 4 (left) we evaluate the ranking in terms of NDCG30@k for several methods. The combination of paired and single attributes is the best performing method. Paired-attributes by itself is better than single attributes and both are better than DeCAF. The performance of the cosine similarity for DeCAF is surprisingly good. Domain adaptation applied on DeCAF features could only slightly improve the ranking for the first top retrieved images. Surprisingly, ITML, a state-of-the-art metric learning approach, degrades the performance of DeCAF, probably because the definition of the metric (i.e. similar/dissimilar pairs) is not straight forward in our case. Classemes, a different attribute-based representation, also could not perform well in this case. These attributes are trained on an external dataset with different feature types, which may explain why they do not reach the performance of our single attributes.

The absolute values of the NDCG30@k metric are small. As explained in section 3.2, this is because we compare the ranking methods with the first 30 ground-truth images only. Increasing the cutoff value leads to higher absolute value of NDCG, however, at the same time, the measure loses its meaning as shown in Fig. 3. Rather than the absolute values, it's the relative performance of different methods that counts. In figure 5, some examples of swap retrieval are shown to assess qualitative performance of attributes versus DeCAF.

**Other classes.** We also evaluate our methods on two other pairs of classes, namely (cow,sheep) and (toilet,chair). In these experiments we have used the same setting but evaluate it on top performing methods only (i.e. DeCAF and attributes). In (cow,sheep) the difference between DeCAF and attributes is smaller than in (cat,dog). One possible explanation is that the scene plays an important role in (cow,sheep) and DeCAF is a strong baseline for capturing scene elements of a picture like grass and sky. On the other hand, attributes are performing much better than DeCAF in the (toilet,chair) pair. Surprisingly single attributes are stronger here probably because the distance between the two classes is higher in this case, so there

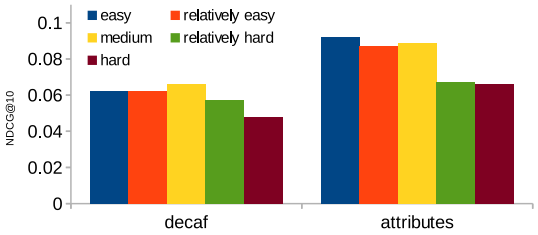


Figure 6: NDCG30@10 for different groups of query for (**left**) DeCAF, (**right**) single+paired attributes

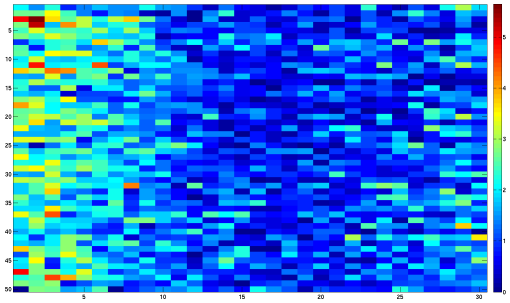

Figure 7: Matrix of agreement between human ranking and text ranking for 50 queries. Each row represents a query. For more details see the text.

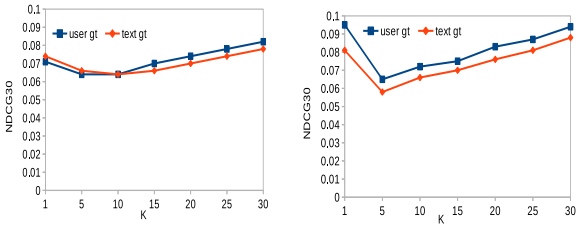


Figure 8: Performance of (**left**) DeCAF, (**right**) single+paired attributes for user ranking vs. text ranking as ground-truth

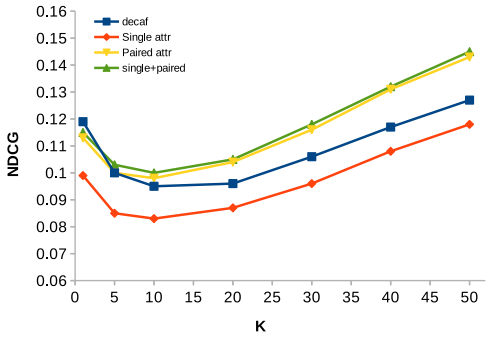


Figure 9: Results of NDCG30@k for standard image retrieval task on COCO dataset.

is not enough class-specific paired-attributes for both classes to train a good paired-attribute detector.

**Query difficulty.** In general, for some queries, there are no relevant images for the swap retrieval task in the dataset. To take this effect into account, we analyze the results in function of the difficulty of the query. We divide the queries in 5 groups based on their difficulty, where the difficulty is measured based on the average similarity of the top 30 most similar images (based on text similarity). As shown in figure 6, there is a direct relation between the quality of a query and its retrieval performance. This trend is more outspoken for the attributes than for DeCAF.

## 5.3 User Evaluation

Here, we analyze the data collected from human evaluations, where human input was used to re-rank the first 30 most similar images for 50 selected queries. To this end, we provide the user a pair of images and ask him/her to select the most similar one to the given query. The $O(n^2)$ growth of comparisons between $n$ images prohibits an exhaustive human evaluation. Based on these annotations, we re-rank the images using ranking SVM and use this as ground truth.

Figure 7 shows for each query (row) how often (relatively speaking) a given image (column) was ranked higher than its competitor in the binary comparisons. Images are ranked from left to right based on textual similarity to the query. The agreement between text and human ranking is quite high. Moreover, human ranking loosely agree that the first 10 images are more similar to the query than the next 20 images. This is consistent with our assumption about the cutoff value. In figure 8 we compare the performance of both DeCAF and attributes for the 50 queries for which the human annotation are available. This figure verifies that the experiments of the previous sections with text as ground-truth are a valid evaluation for this task.

## 5.4 Image Retrieval

Finally, in order to assess the usefulness of our attributes in a normal retrieval task, we set up an experiment by selecting training and validation data from 80 categories of the COCO dataset. The task is to find similar object categories to the query image. Same as above, the validation part of the dataset is used for testing in a leave-one-query-out strategy. We sub-sample 200 images from each of these 80 categories for our experiments. To be consistent with the previous experiments, the **NDCG30** is used for reporting results. For finding single attributes, 10 discriminative words from each pair of classes are selected (Eq. 1) and then aggregated to build one list of candidates for the dataset. We use the same method as explained in section 4.5 for extracting paired-attributes except that in this case we only use common paired-attributes and there is no symmetric class-specific paired-attributes.

As shown in figure 9, the combination of single and paired attributes are outperforming DeCAF in $k > 5$ which suggests the effectiveness of the proposed attributes in a standard retrieval task.

# 6. CONCLUSION

In this work we have presented swap retrieval, a new and challenging task that combines the semantic of text-based queries and the power of example-based retrieval. Given an image query containing a certain object class, it consists of searching for an image with similar content but swapping the object class with another similar one.

We have adapted a metric used in standard retrieval to the proposed swap retrieval and through a human validation we have shown that the text associated to images can be used for meaningful evalu-

ations. In this sense we have compared standard methods for image retrieval and image classification and have shown that their performance on this task is low. Instead, a better performance is obtained by a set of learned attributes that can better represent the semantics of an image and therefore it is better suited for swap retrieval.

## 7. ACKNOWLEDGMENT

The work was supported in part by FWO through the project G.0.398.11.N.10 "Multi-camera human behavior monitoring and unusual event detection" and FP7 ERC Grant 240530 COGNIMUND.

## 8. REFERENCES

[1] A. Andoni and P. Indyk. Near-optimal hashing algorithms for approximate nearest neighbor in high dimensions. *Commun. ACM*, 51(1):117–122, 2008.

[2] T. L. Berg, A. C. Berg, and J. Shih. Automatic attribute discovery and characterization from noisy web data. In *ECCV*, 2010.

[3] R. Datta, D. Joshi, J. Li, and J. Z. Wang. Image retrieval: Ideas, influences, and trends of the new age. *ACM Comput. Surv.*, 40(2), 2008.

[4] J. V. Davis, B. Kulis, P. Jain, S. Sra, and I. S. Dhillon. Information-theoretic metric learning. In *ICML*, 2007.

[5] J. Donahue, Y. Jia, O. Vinyals, J. Hoffman, N. Zhang, E. Tzeng, and T. Darrell. Decaf: A deep convolutional activation feature for generic visual recognition. *arXiv preprint arXiv:1310.1531*, 2013.

[6] M. Douze, A. Ramisa, and C. Schmid. Combining attributes and fisher vectors for efficient image retrieval. In *CVPR*, 2011.

[7] B. Fernando, A. Habrard, M. Sebban, and T. Tuytelaars. Unsupervised visual domain adaptation using subspace alignment. In *ICCV*, 2013.

[8] Y. Gong and S. Lazebnik. Iterative quantization: A procrustean approach to learning binary codes. In *CVPR*, 2011.

[9] A. Gordo, J. A. Rodríguez-Serrano, F. Perronnin, and E. Valveny. Leveraging category-level labels for instance-level image retrieval. In *CVPR*, 2012.

[10] D. Grangier and S. Bengio. A discriminative kernel-based approach to rank images from text queries. *TPAMI*, 30(8):1371–1384, 2008.

[11] A. Habibian, T. Mensink, and C. G. M. Snoek. Composite concept discovery for zero-shot video event detection. In *ICMR*, 2014.

[12] H. Jégou, M. Douze, and C. Schmid. Product quantization for nearest neighbor search. *TPAMI*, 33(1):117–128, 2011.

[13] A. Kovashka and K. Grauman. Attribute pivots for guiding relevance feedback in image search. In *ICCV*, 2013.

[14] A. Kovashka, D. Parikh, and K. Grauman. Whittlesearch: Image search with relative attribute feedback. In *CVPR*, 2012.

[15] N. Kumar, A. C. Berg, P. N. Belhumeur, and S. K. Nayar. Describable visual attributes for face verification and image search. *TPAMI*, 33(10):1962–1977, 2011.

[16] R. Layne, T. M. Hospedales, and S. Gong. Re-id: Hunting attributes in the wild. In *BMVC*, 2014.

[17] C. Li, D. Parikh, and T. Chen. Automatic discovery of groups of objects for scene understanding. In *CVPR*, 2012.

[18] X. Li, C. G. M. Snoek, M. Worring, and A. W. M. Smeulders. Harvesting social images for bi-concept search. *IEEE Transactions on Multimedia*, 14(4):1091–1104, 2012.

[19] T.-Y. Lin, M. Maire, S. Belongie, L. Bourdev, J. Hays, P. Perona, D. Ramanan, P. Dollar, and L. Zitnick. Microsoft coco: Common objects in context. In *ECCV*, 2014.

[20] P. Over, G. Awad, M. Michel, J. G. Fiscus, G. Sanders, B. Shaw, W. Kraaij, A. F. Smeaton, and G. Quénot. TRECVID 2012 - an overview of the goals, tasks, data, evaluation mechanisms and metrics. In *Proceedings of the TRECVID Conference*, 2012.

[21] D. Parikh and K. Grauman. Relative attributes. In *ICCV*, 2011.

[22] J. C. Pereira and N. Vasconcelos. Cross-modal domain adaptation for text-based regularization of image semantics in image retrieval systems. *CVIU*, 124:123–135, 2014.

[23] J. Philbin, O. Chum, M. Isard, J. Sivic, and A. Zisserman. Lost in quantization: Improving particular object retrieval in large scale image databases. In *CVPR*, 2008.

[24] D. Qin, C. Wengert, and L. J. V. Gool. Query adaptive similarity for large scale object retrieval. In *CVPR*, 2013.

[25] A. Quattoni, M. Collins, and T. Darrell. Learning visual representations using images with captions. In *CVPR*, 2007.

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

[27] K. Saenko, B. Kulis, M. Fritz, and T. Darrell. Adapting visual category models to new domains. In *ECCV*, 2010.

[28] R. Salakhutdinov and G. E. Hinton. Semantic hashing. *Int. J. Approx. Reasoning*, 50(7):969–978, 2009.

[29] A. Sharif Razavian, H. Azizpour, J. Sullivan, and S. Carlsson. Cnn features off-the-shelf: An astounding baseline for recognition. In *CVPR Workshops*, 2014.

[30] N. V. Shirahatti and K. Barnard. Evaluating image retrieval. In *CVPR (1)*, 2005.

[31] B. Siddiquie, R. S. Feris, and L. S. Davis. Image ranking and retrieval based on multi-attribute queries. In *CVPR*, 2011.

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

[33] J. Wang, S. Kumar, and S. Chang. Semi-supervised hashing for large-scale search. *TPAMI*, 34(12):2393–2406, 2012.

[34] Y. Weiss, A. Torralba, and R. Fergus. Spectral hashing. In *NIPS*, 2008.

[35] F. X. Yu, L. Cao, R. S. Feris, J. R. Smith, and S. Chang. Designing category-level attributes for discriminative visual recognition. In *CVPR*, 2013.

[36] F. X. Yu, R. Ji, M.-H. Tsai, G. Ye, and S.-F. Chang. Weak attributes for large-scale image retrieval. In *CVPR*, 2012.

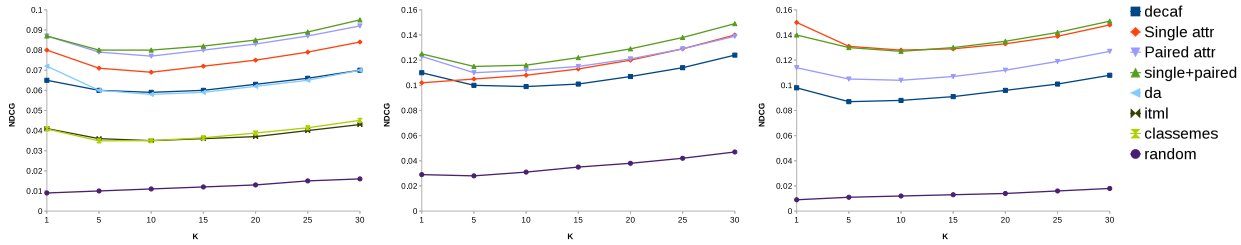


Figure 4: NDCG30@k for different methods using label information for (**left**) cat and dog (**middle**) cow and sheep and (**right**) chair and toilet


Figure 5: Qualitative examples of swap retrieval. The first column is a query. The next three are the retrieved images using single and paired attributes and next three are the retrieved images using DeCAF. The last row is an example of poor retrieval based on attributes.