Distributional Semantics with Eyes: Using Image Analysis to Improve Computational Representations of Word Meaning

Elia Bruni
CIMeC, University o fTrento
elia.bruni@unitn.it

Jasper Uijlings
DISI, University of Trento
jrr@disi.unitn.it

Marco Baroni
CIMeC, University of Trento
marco.baroni@unitn.it

Nicu Sebe
DISI, University of Trento
sebe@disi.unitn.it

ABSTRACT

The current trend in image analysis and multimedia is to use information extracted from text and text processing techniques to help vision-related tasks, such as automated image annotation and generating semantically rich descriptions of images. In this work, we claim that image analysis techniques can "return the favor" to the text processing community and be successfully used for a general-purpose representation of word meaning. We provide evidence that simple low-level visual features can enrich the semantic representation of word meaning with information that cannot be extracted from text alone, leading to improvement in the core task of estimating degrees of semantic relatedness between words, as well as providing a new, perceptually-enhanced angle on word semantics. Additionally, we show how distinguishing between a concept and its context in images can improve the quality of the word meaning representations extracted from images.

Categories and Subject Descriptors

H.4 [**Information Systems Applications**]: Miscellaneous

General Terms

Algorithms, Experimentation, Theory

Keywords

Semantics, language, visual words, object recognition

1. INTRODUCTION

In recent years, the image analysis community has shown much interest in using information extracted from text and techniques from text processing to tackle vision-related tasks, such as automated image annotation, generating textual descriptions of images, region naming or exploiting textual information to induce visual attributes [2, 7, 20, 21, 25]. In this work, we take a bold step in the opposite direction: we claim that image analysis techniques are mature enough that we can use features automatically extracted from images as a surrogate for a general-purpose representation of word meaning in semantic text analysis.

Our approach is based on the widely adopted *distributional hypothesis*, stating that words that are semantically similar occur in similar contexts [36]. Concretely, the distributional hypothesis is implemented in computational linguistics by the methods of *distributional semantics*, that approximate word meanings with vectors that keep track of patterns of co-occurrence of words in large collections of text [27, 49, 53]. Extended empirical evidence has confirmed that distributional semantics is very good at harvesting effective meaning representations on a large scale, because words that have similar meanings tend to occur in similar textual contexts, and thus co-occur with the same context words. For example, both *car* and *automobile* occur with terms such as *street*, *gas* and *driver*, and thus their distributional vectors will be very similar, cuing the fact that these words are synonyms. Distributional semantic vectors can be used in pretty much all applications that require a representation of word meaning, ranging from document retrieval and classification to question answering, automated thesaurus construction and machine translation [53].

Despite its success, distributional semantics is severely limited by the fact that, in the aforementioned approach, the meaning of words is entirely given by other words. Besides the obvious difference with the way in which humans acquire meaning (not only from language, but also from perception), this limitation leads to empirical weaknesses. For example, many studies [1, 4, 5, 46] have underlined how text-derived distributional models capture encyclopedic and functional properties of word meanings, but miss their concrete aspects. Intuitively, we might harvest from text the information that *bananas* are *tropical* and *eatable*, but not that they are *yellow* (because few authors feel the need to write down obvious statements such as "*bananas are yellow*").

The novel idea of this work is to enrich text-based distributional semantic vectors with features automatically extracted from images that are associated with words in tagged

MM'12, October 29–November 2, 2012, Nara, Japan.

image collections. Just like co-occurrence counts of words with other words in text documents are used to populate text-based distributional vectors, we use counts of quantized SIFT-based *visual words* [51] across all the images that are tagged with a (textual) word in an image collection to populate image-based distributional semantic vectors. The procedure is unsupervised and relatively knowledge-lean (requiring only a possibly noisy collection of labeled images).

This study presents evidence that simple, relatively low-level visual features such as visual words can indeed enrich the semantic representation of word meaning with information that is missing when these models are developed from text alone, leading to improvement in the core task of estimating degrees of semantic relatedness between words, as well as providing a new, perceptually-enhanced angle on word semantics. We explore moreover different ways in which text- and image-based vectors can be combined to integrate the two information sources. Finally, we present a preliminary experiment in which we revisit the distributional hypothesis from the point of view of image analysis, showing how distinguishing between a concept and its context in images can improve the quality of the semantic vectors extracted from the images.

2. RELATED WORK

Nowadays huge image collections are freely available on the Web, often incorporating additional textual data such as tags, which provide complementary information related to the image content. The multimedia and computer vision communities have fruitfully used this information to supplement image analysis and to help bridging the semantic gap that visual features cannot fill on their own. Taking inspiration from methods originally used in text processing, algorithms for image labeling, search and retrieval have been built upon the connection between text and visual features.

Barnard *et al.* [2] present one of the first attempts to model multimodal sets of images with associated text, learning the joint distribution of image regions and concrete concepts. Recently this has been extended to attributes such as *yellow* or *striped* [7, 20, 26, 59], enabling transfer learning [20, 26] and even unseen object recognition [26]. Rohrbach *et al.* [47] enhance transfer learning for attribute-based classification by using semantic relatedness values that they extract from linguistic knowledge bases.

The works reviewed above focus mostly on the visual domain where, except for [7], most use keywords rather than natural language captions. Both Farhadi *et al.* [21] and Kulkarni *et al.* [25] aim to create more natural descriptions for images than just tags. They first use visual features to predict the content of an image in terms of objects and attributes. Then they use a natural language generation system to create image captions.

Zha *et al.* [61] presented a system for visual query suggestion for image search. When a user types a query, they do not only suggest queries that are semantically similar, but also provide top-ranked images for all suggested queries to aid the user in their search.

Many papers address the topic of multimodal media retrieval (see, e.g., [16, 52]), where the fusion of different media is an important problem. This can be done with early fusion (concatenating feature vectors containing information from different modalities), e.g., [43, 13], late fusion (concatenating outputs of single-modality systems), e.g., [17, 58], or using a joint model such as with Canonical Correlation Analysis, e.g., [29, 45]. In this paper we use early fusion, possibly followed by latent modality smoothing.

In all aforementioned efforts the goal was to use text to improve image-related tasks, and typically they attempt to model the relation between specific images and specific textual passages. In contrast, (i) we want to use image-derived features to improve the representation of word meaning and (ii) we are interested in modeling the meaning of word *types* on the basis of sets of images connected to a word, and not to model specific word-image relations.

Closer to our goals, Barnard and Johnson [3] use information from combined text and images to improve word sense disambiguation. However, their purpose is to distinguish between senses of a word in a specific context and not to create general-purpose semantic representations. Additionally, their method necessitates a sense-disambiguated corpus of images. Bergsma and Goebel [8] use image-extracted features to train a classifier to decide whether a noun is a plausible object of a verb. This could be a specific application of our more general approach. Similarly, in [11] we use an earlier version of the system presented here (featuring a less performing visual word extraction method and no SVD-based smoothing) to study the semantics of color adjectives.

Early attempts to implement the same idea we present here were made in [23, 31] and in our own prototype work in [12]. None of these earlier studies goes beyond proof-of-concept evidence. We introduce here a fully fleshed-out model of image-based distributional vector extraction and image-text fusion. We report, for the first time, multimodal performance significantly above that of a textual model at the state of the art, on much larger test sets than those used in the earlier experiments. We present moreover an extensive qualitative analysis of the semantic properties of image-based semantic vectors. Finally, we are the first to evaluate the impact that localizing an object in an image has on the image-based semantic representation of the meaning of the corresponding word.

3. SEMANTIC VECTORS

3.1 Text-based semantic vectors

Text-based distributional semantic models approximate the meaning of words with vectors that record their distributional history in a corpus [53] (see [33] for a formal treatment). A distributional semantic model is encoded in a matrix whose m rows are **semantic vectors** representing the meanings of a set of m **target words**. Each component of a semantic vector is a function of the occurrence counts of the corresponding target word in a certain context. Definitions of context range from simple ones (such as documents or the occurrence of another word inside a fixed window from the target word) to more linguistically sophisticated ones (such as the occurrence of certain words connected to the target by special syntactic relations) [41, 48, 53]. After the raw target-context counts are collected, they are transformed into **association scores** that typically discount the weights of components whose corresponding word and context pairs have a high probability of chance co-occurrence [19]. The rank of the matrix containing the semantic vectors as rows can optionally be decreased by **dimensionality reduction**, that might provide beneficial smoothing by getting rid of noise components and/or allow more efficient storage and computation [9, 27, 48, 50]. Finally, the distributional semantic similarity of a pair of target words is estimated by a **similarity function** that takes their semantic vectors as input and returns a scalar similarity score as output.

In this study we harvest **text-based semantic vectors** from the freely available ukWaC and Wackypedia corpora (about 3 billion words in total)[1]. Since both corpora have been automatically annotated with lemma and part-of-speech information, we take both into account when extracting target and context words (e.g., the string "*sang*" is treated as an instance of the *verb lemma* "*sing*"). We collect semantic vectors for a set of 30K target words (lemmas), namely the top 20K most frequent nouns, 5K most frequent adjectives and 5K most frequent verbs in the combined corpora. The same 30K lemmas are also employed as contextual elements (consequently, our text-based semantic model is encoded in a 30K×30K matrix).

We adopt a relatively simple definition of context, in terms of words that co-occur within a window of fixed width, in the tradition of the popular HAL model [34]. Window-based models have been reported to be at the state of the art in various semantic tasks [44, 49], and in [11] we show that the window-based method we use here outperforms both document- and syntax-based models on the semantic benchmarks introduced in Section 4.2 below. The vectors record the co-occurrence of targets with context words within windows of 20 context items to the left and right of the targets (similar performance is obtained with a narrower window of 2 words left and right).

We transform raw co-occurrence counts into nonnegative Local Mutual Information (**LMI**) association scores. LMI scores are obtained by multiplying raw counts by Pointwise Mutual Information, and in the nonnegative case they are a close approximation to the popular Log-Likelihood Ratio scores [19]. The nonnegative LMI of target t and context c is defined as:

$$LMI(t,c) = \max\left(\mathrm{Count}(t,c) \times \log \frac{\mathrm{P}(t,c)}{\mathrm{P}(t)\mathrm{P}(c)}, 0\right)$$

We do not apply dimensionality reduction to the text-based distributional matrix. However, we will use it as a text and image information mixing technique in Section 4.2 below. For these purposes, we adopt the Singular Value Decomposition (**SVD**), a widely used method to find the best approximation of the original data points in a space of lower underlying dimensionality whose basis vectors ("principal components" or "latent dimensions") are selected to capture as much of the variance in the original space as possible [35, Ch. 18]. Following the description in [43], the SVD of a matrix M of rank r is a factorization of the form

$$M = U\Sigma V^t \tag{1}$$

where

$$\begin{cases} U : \text{matrix of eigenvectors derived from } MM^t \\ \Sigma : r \times r \text{ diagonal matrix of singular values } \sigma \\ \sigma : \text{square roots of the eigenvalues of } MM^t \\ V^t : \text{matrix of eigenvectors derived from } M^tM \end{cases}$$

Finally, our similarity function is the **cosine** of the angle formed by two semantic vectors. The cosine is by far the most common similarity measure in the distributional semantics literature, and the soundest one given the geometric approach adopted by these models. The cosine of two semantic vectors **a** and **b** is their dot product divided by the product of their lengths:

$$\cos(\mathbf{a},\mathbf{b}) = \frac{\sum_{i=1}^{i=n} a_i \times b_i}{\sqrt{\sum_{i=1}^{i=n} a_i^2} \times \sqrt{\sum_{i=1}^{i=n} b_i^2}}$$

The cosine ranges from 0 (orthogonal vectors) to |1| (parallel vectors pointing in the same or opposite directions). When all vector components are nonnegative, the cosine is also nonnegative.

3.2 Image-based semantic vectors

We hypothesize that, given a set of target words, semantic vectors can be extracted from a corpus of images associated with the target words using a similar pipeline to what is commonly used to construct text-based vectors: Collect co-occurrence counts of target words and image-based contexts, transform them into association weights, possibly reduce dimensionality, and finally approximate the semantic relatedness of two target words by a similarity function over the vectors representing them. In particular, we use regular bag-of-visual-words representations of images [14, 40, 51, 54, 60] and consequently our **image-based semantic vectors** count how many times each visual word occurs across all images that contain a specific concept. In the first two experiments we take all visual words of each image tagged with the concept. In the final experiment all concepts denote objects and we capture visual words from the context and object separately.

ESP image vectors. In Section 4.1 and 4.2 we extract image-based semantic vectors from the ESP-Game data set[2], containing 100K images labeled through a game with a purpose (like the text data, these images are a random sample from the Web, and they are not chosen to be representative of the words they are labeled with). There are 20K distinct word tags in the dataset, with an average of 4 words per image. Consequently, we can build image-based semantic vectors for these 20K target words. For comparability with the text-based models, the targets are lemmatized and annotated with parts of speech using a heuristic method.

Using the publicly available VLFeat library[3], we extract HSV-based SIFT descriptors [10, 32] on a regular grid with five pixels spacing, at four multiple scales (10, 15, 20, 25 pixel radii) and zeroing the low contrast descriptors. To map SIFT descriptors to visual words, we cluster the descriptors in their 128-dimensional space using the k-means clustering algorithm. We set k, the size of the visual word vocabulary, to 5,000. Some spatial information is preserved using pyramid spatial representations [24, 30]. The image is divided into regions, each region is represented as a bag-of-visual-words (BoVW) vector, and then these vectors are concatenated. For the ESP dataset, the regions are obtained by dividing the image in 4×4, for a total of 16 regions. The resulting image-based semantic vectors therefore have a dimensionality of $5,000 \times 16 = 80,000$. The parameters of the visual word extraction pipeline are standard choices in the image analysis literature and were adopted without tuning.

[1] http://wacky.sslmit.unibo.it

[2] http://www.espgame.org

[3] http://www.vlfeat.org

Once the BoVW representations are built, each target (textual) word is associated to the list of images which are labeled with it; the visual word occurrences across the list of images is summed to obtain the co-occurrence counts associated with the target (textual) word. Just like in text, raw counts are transformed into LMI scores.

Pascal VOC image vectors. In Section 4.3 we extract image-based semantic vectors from the Pascal VOC 2007 dataset [18], a widely used dataset in computer vision and multimedia with 5011 training images and 4952 test images, containing a total of 20 concepts divided in the categories *animal, vehicle*, and *indoor. Person* is considered a separate category. All images are annotated using bounding boxes outlining all instances of each object (this is the reason why we shift to this dataset for the current experiment). An image may contain multiple instances of the same or different classes.

In the experiment of Section 4.3 we want to distinguish between context and object, which we not only do using ground-truth annotation but also using a BoVW localization framework. The localization framework requires pixel-wise sampled visual words, which we then also use to generate our image-based semantic vectors. Hence this section uses a different, yet also common, BoVW representation.

Specifically, we sample SIFT [32], HSV-SIFT as above [10] and RGB-SIFT [55] at every pixel at a single scale of 16 by 16 pixels, using the fast, publicly available code of [54]. Visual vocabularies are created with a Random Forest [38], using four trees of depth ten, resulting in 4,096 visual words per SIFT variant. Image-based semantic vectors are then created by considering all images containing a specific object (defined by the global, ground truth labeling), using visual words within the following three image regions: (i) the whole image, (ii) only the object region, (iii) only the context surrounding the object region. A visual word is considered object if its centre is contained within the bounding box denoting its location. The location is either given by the ground truth or by the object localization framework.

To automatically generate object locations, we created an object localization framework after [56], using their publicly available selective search code. Selective search uses a variety of hierarchical segmentations to generate a limited set of potential object locations, drastically reducing the number of locations to consider compared to the more common exhaustive search (e.g., [15, 22, 57]). For localization, initial object models are trained using ground-truth examples of target and non-target objects, as well as locations generated by selective search that partially overlap with the target object but which do not capture it accurately. Specifically, these have an overlap score between 20% and 50% according to the Pascal overlap criterion [18]. In the testing phase the object models are applied to all locations of each image generated by selective search. All object models are retrained using hard negative examples [22, 28], which are obtained by applying the models on the training set and collecting the highest scored locations of negative training images.

4. EXPERIMENTS

Our first experiment (Section 4.1) demonstrates that image-based semantic vectors, just like text-based ones, encodes meaningful semantic information about the words they represent, and provides qualitative evidence that vectors derived from the two sources differ in interesting ways. Section 4.2, then, compares the performance of text-based and image-based models quantitatively on two semantic benchmarks, and explores two ways to combine them, evaluated on the same benchmarks. We show that, while image-based vectors still lag behind text-based vectors, their combination significantly outperforms text alone. Finally, Section 4.3 is a pilot study in which we reconsider the notion of context in image-based models, showing that recognizing the object denoted by a word in an image, thus taking take the object-surround distinction into account, can improve performance of image-based vectors on a semantic task.

4.1 Semantic properties of image-based word representations

A few decades of experimentation have demonstrated that vectors of shallow text-extracted features succeed in capturing many aspects of word meaning. We want to show, first of all, that word-representing vectors constructed from simple bag-of-visual-words features extracted from the images that co-occur with the target words are also semantically meaningful. For this purpose, we use the **BLESS** data set introduced in [6].

4.1.1 Dataset and method

BLESS contains a set of 200 **pivot** words denoting concrete concepts (we use 184 pivots, since for the remaining 16 we did not have a sufficiently large set of related words covered by our models). For each of the pivots, the data set contains a number of related words, or **relata**, instantiating the following 8 **semantic relations** with the pivots: COORD: the relatum is a noun that is a co-hyponym (coordinate) of the pivot (*alligator-lizard*); HYPER: the relatum is a noun that is a hypernym (superordinate) of the pivot (*alligator-reptile*); MERO: the relatum is a noun referring to a meronym, that is, a part or material of the pivot (*alligator-teeth*); ATTRI: the relatum is an adjective expressing an attribute of the pivot (*alligator-aquatic*); EVENT: the relatum is a verb referring to an action or event involving the concept (*alligator-swim*); RAN.N, RAN.J and RAN.V, finally, are control cases where the pivot is matched to a set of random nouns (*alligator-trombone*), adjectives (*alligator-electronic*) and verbs (*alligator-conclude*), respectively. These random relata represent semantically unrelated words.

For each pivot, BLESS contains a set of relata of each category (ranging from 7 hypernyms to 33 random nouns per pivot on average). In this way, BLESS can highlight the broader semantic properties of a model independently of its more specific preferences. For example, both a model that assigns a high score to *alligator-aquatic* and a model that assigns a high score to *alligator-green* will be correctly treated as models that have picked a relevant attribute of *alligators*. At the same time, the comparison of the specific relata selected by the models allows a more granular qualitative analysis of their differences.

Following the guidelines of [6], we analyze a semantic model as follows. We compute the cosine between the model vectors representing each of the 184 pivots and each of its relata, picking the relatum with the highest cosine for each of the 8 relations (the nearest hypernym, the nearest random noun, etc.). We then transform the 8 similarity scores collected in this way for each pivot onto standardized z scores (mean = 0, std = 1, to get rid of pivot-specific effects), and produce a boxplot summarizing the distribution of scores

pivot	*text*	*image*	*pivot*	*text*	*image*
cabbage	leafy	white	helicopter	heavy	old
carrot	fresh	orange	onion	fresh	white
cherry	ripe	red	oven	electric	new
deer	wild	brown	plum	juicy	red
dishwasher	electric	white	sofa	comfortable	old
elephant	wild	white	sparrow	wild	little
glider	heavy	white	stove	electric	hot
gorilla	wild	black	tanker	heavy	grey
hat	white	old	toaster	electric	new
hatchet	sharp	short	trout	fresh	old

Table 1: Attributes preferred by text- vs. image-based models

per relation across the 184 pivots (for example, the leftmost box in the first panel of Fig. 1 reports the distribution of 184 standardized cosines of nearest coordinate relata with the respective pivots).

In this experiment, the image-based model is trained on the ESP-Game data (see Section 3.2 above).

4.1.2 Results

In Fig. 1, we report BLESS nearest relata distributions for our text- and image-based models. The patterns produced by the text-based model (left panel) illustrate how a sensible word meaning profile should look like: coordinates are the most similar terms (an *alligator* is maximally similar to a *crocodile*), followed by superordinates (*reptile*) and parts (*teeth*). Semantically related adjectives (ATTRI: *aquatic*) and verbs (EVENT: *swim*) are less close to the pivots, but still more so than any random item.

The right panel shows the distribution of relata in the image-based semantic vectors. The overall pattern is quite similar to the one observed with the text-based vectors: There is a clear preference for coordinates, followed by hypernyms and parts, then attributes and events, with all random relata further away from the pivots than the semantically meaningful categories. For both models, coordinates are significantly closer to pivots than hypernyms and meronyms, that are significantly closer than attributes and events, that are in turn significantly closer than any random category (Tukey Honestly Significant Difference tests, $p \leq 0.05$). Although the difference between hypernyms and parts is not significant with either representation, intriguingly the image-based vectors show a slight preference for the more imageable parts (*teeth*) than the more abstract hypernyms (*reptile*). The only difference of statistical import is the one between events and attributes, where the text-based model shows a significant preference for events, whereas the two categories are statistically indistinguishable in the image-based model (as we will see shortly, the relative preference of the latter for attributes is probably due to its tendency to pick perceptual adjectives denoting color and size).

Looking more closely at the specific relata picked by the text- and image-based models, the most striking differences pertain, again, to attributes. The text- and image-based models picked the same attribute for a pivot in just 20% of the cases (compare to 40% overlap across all non-random relation types). Table 1 reports the attributes picked by the text- vs. image-based models for 20 random cases where the two mismatch.

It is immediately clear from the table that, despite the fact that the pivots are nouns denoting concrete concepts, the text-based model almost never picks adjectives denoting salient perceptual properties (and in particular visual properties: just *white* for *hat* and *leafy* for *cabbage*). The text-based model focuses instead on encyclopedic properties such as *fresh, ripe, wild, electric* and *comfortable.* This is in line with earlier analyses of the "ungrounded" semantics provided by text-based models [1, 4, 5, 46], and differs greatly from the trend encountered with the image-based model. In 12/20 cases, the closest attribute for the latter is a color. In the remaining cases, we have size (*short*, *little*), one instance of *hot* and, surprisingly, four of *old.*

To conclude, the analysis we presented confirms, on the one hand, our hypothesis that image-based distributional vectors contain sufficient information to capture a network of sensible word meaning relations. On the other, there are intriguing differences in the relations picked by the text- and image-based models, pointing to their complementarity. We turn now to ways in which the two sources of information can be combined to obtain better representations of word meaning.

4.2 Evaluating and integrating text- and image-based representations of word meaning

Given that word meaning representations extracted from images are not only informative, but also of a complementary nature with respect to text-based representations, we next quantify the performance of both text- and image-based models (**Text** and **Image** in Table 2 below) on two semantic relatedness benchmarks, and we explore two ways in which they can be combined (in both cases, only vectors for the 20K target words for which we have both text- and image-based representations are considered).

The first combination method, that we call **Linear**, is to simply row-normalize, linearly weight and concatenate the text- and image-based vectors, similarly to the approach we originally presented in [12]. The other strategy, that we call **Smoothed**, consists in projecting concatenated text- and image-based vectors onto a lower dimensionality latent space using SVD, in order to promote the formation of new connections within the components from each modality taking into account information and connections present in the other modality (see [13] for similar ideas applied to image annotation and retrieval tasks). SVD is applied to the matrix obtained by row-normalizing and then concatenating the text- and image-based vectors representing the target words. In the SVD factorization, we set all but the top k singular values to 0, obtaining a matrix with the original size but lower rank. Again, after SVD, we linearly combine the text- and image-based blocks of the reduced matrix. Essentially, Smoothed is identical to Linear, except for the SVD smoothing step.

4.2.1 Datasets and method

As is standard in the distributional semantic literature [49, 53], we assess the performance of our models on the task of predicting the degree of semantic relatedness between two words as rated by human judges. The many applications of automated measures of word relatedness include word sense disambiguation, query expansion, textual advertising and information extraction.

More specifically, we test the distributional models on the

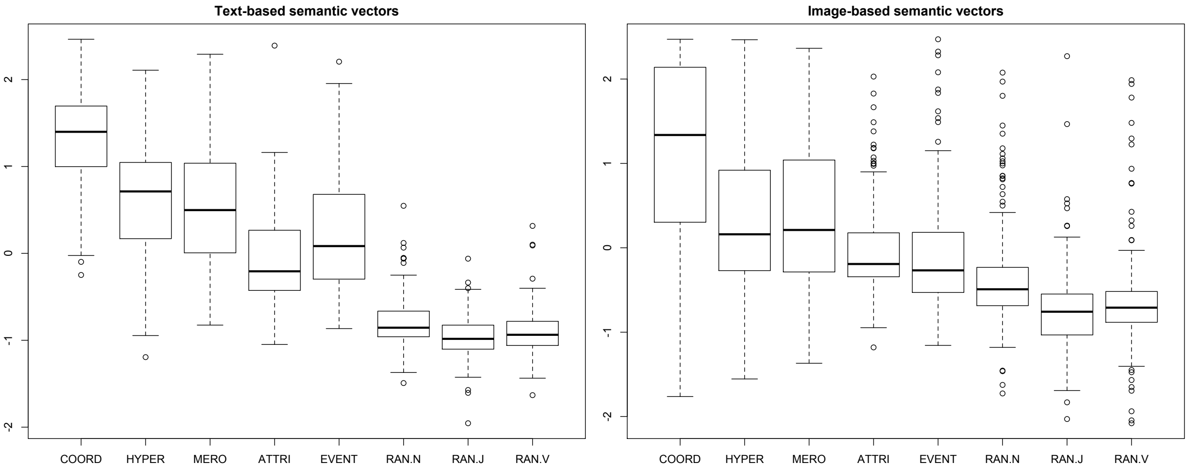

Figure 1: Distribution of z-normalized cosines of words instantiating various relations across BLESS pivots

MEN and **WS** benchmarks. WS, that is, WordSim353[4], was constructed by asking 13 subjects to rate a set of 353 word pairs on an 11-point meaning similarity scale and averaging their ratings (e.g., *dollar/buck* has a very high average rating, *professor/cucumber* a very low one). Our target words cover 252 WS pairs (thus, the correlations reported below are not directly comparable to those reported in other studies that used WS).

MEN[5] consists of 3,000 word pairs with $[0, 1]$-normalized semantic relatedness ratings provided by Amazon Mechanical Turk workers (at least 50 ratings per pair). For example, *beach/sand* has a MEN score of 0.96, *bakery/zebra* received a 0 score. Our target words cover all MEN pairs. Following [11], we use 2,000 MEN pairs for model tuning and 1,000 pairs for evaluation.

Models are evaluated as follows. For each pair in a dataset, we compute the cosine of the model vectors representing the words in the pair, and then calculate the Spearman correlation of these cosines with the (pooled) human ratings of the same pairs.

The Linear model has a weight parameter $0 \leq \alpha \leq 1$ determining the relative importance given to text- and image-based components ($\mathbf{combined} = \alpha\mathbf{text} + (1 - \alpha)\mathbf{image}$). The optimal α value found on the MEN development set is 0.4. The Smoothed model has an extra parameter k (number of preserved singular values). Optimal settings on the development data are $\alpha = 0.6$ and $k = 1024$.

In this experiment, as in the previous one, the image-based model is trained on the ESP-Game data (Section 3.2).

4.2.2 Results

Table 2 reports the correlations of the models on the MEN (1,000 pairs not used for parameter tuning) and WS datasets. The first two rows show the separate results for the text- and image-based models. Text shows good performance in both datasets. The Image cosines are significantly correlated with human ratings in both datasets (confirming the promising BLESS results), but correlations are considerably below those attained by Text. The next two rows show the correlations of the combined models. The Linear approach outperforms Text on MEN but not on WS. The Smoothed combination, on the other hand, outperforms Text on both benchmarks (with the improvement significant at $p < 0.001$ on MEN, according to a a two-tailed paired permutation test [37]).[6]

Model	*MEN*	*WS*
Text	0.68	0.70
Image	0.43	0.36
Linear	0.73	0.67
Smoothed	**0.76**	**0.75**

Table 2: Spearman correlation of the models on MEN and WS (all coefficients significant with $p <$ 0.001).

The main results of the semantic benchmark tests reported in this section are (i) that image-based semantic vectors significantly correlate with human semantic intuition about word meaning (albeit to a lesser degree than text-based vectors); and (ii) that combining the two information sources is beneficial, but not trivial (the Smoothed approach significantly outperforms Text, the simple Linear combination not quite).

4.3 The illustrated distributional hypothesis

Recall from the introduction that distributional semantics is grounded in the distributional hypothesis that words that are semantically similar occur in similar contexts [36]. Our extension to visual contexts until here has been straightforward: If a target word is associated to a set of images, then all visual features extracted from the images count as the context in which the word occurs. However, if a word is used to tag an image, the concept denoted by the word will typically be present in the image (at least if it is a concrete concept). We can thus think of a more thorough application of the distributional hypothesis to images: Semantically

[4] `http://www.cs.technion.ac.il/~gabr/resources/data/wordsim353`

[5] `http://clic.cimec.unitn.it/~elia.bruni/MEN`

[6] In experiments not reported here, we verified that the Smoothed model outperforms Text even when the latter undergoes SVD, thus the improvements brought about by Smoothed cannot be attributed to SVD alone.

similar objects will tend to occur in similar environments in images.

To verify this hypothesis, we use three different representations of the images associated with a target word. (i) The common, global bag-of-words representation which captures the appearance of the whole image, the de-facto representation for visual recognition, and the same representation we used in all experiments above (**Global** approach in the discussion below). (ii) A bag-of-words representation of the object only (**Obj**). (iii) A bag-of-words representation of the context (or surround) only (**Surr**). This latter representation would be the counterpart of the textual context representation, where a target word is represented by the words that surround it within a window. In the current experiment we create a distinction between object and context using both ground truth object locations as a theoretical experiment, as well as locations resulting from an object localization framework. The ground truth division is denoted **GT**, the automatic localization is denoted **AL**.

4.3.1 Dataset and method

We test the models by measuring their correlation to human judgements on word similarity. We created a new evaluation benchmark for this purpose as follows. We first formed every possible pairing of the 20 Pascal concept words (see Section 3.2 above), obtaining 190 pairs in total. Then we obtained relatedness ratings for the pairs by crowdsourcing using Amazon Mechanical Turk. We presented Turkers with a list of two candidate word pairs, each pair randomly matched with a comparison pair sampled without replacement from the same list and rated in this setting (as either more or less related than the comparison point) by a single Turker (thus, we have no measure of inter-rater agreement). In total, each pair was rated in this way against 50 comparison pairs, thus obtaining a final score on a 50-point scale (then normalized between 0 and 1), although the Turkers had to make simple binary choices, that were preferred over an absolute rating for two reasons. First, judgements are often more objective in binary comparisons: for example, there is more agreement that the pair <*cat,dog*> is more similar than <*cat,cow*> than on rating the similarity of <*cat,cow*> on a 10-point scale [42, 39]. Secondly, binary choices allowed the elimination of bad Turkers trough a few straightforward control comparisons between word pairs which we judged ourselves beforehand.

In this experiment, image-based semantic vectors are extracted from the VOC dataset described in Section 3.2. See that section also for details on the automated localization algorithm. We use semantic representations extracted from test images only. The automatic object locations on the test set are generated by a localization framework which we trained on the training set.

4.3.2 Results

Table 3 reports the correlations of our models with human similarity ratings (p<0.0001 for all correlations). The model based on the global approach (Global) already achieves a good correlation with human judgements. Furthermore, we see that there is little difference in performance between the methods based on ground-truth locations and the ones based on automatically generated locations, which proves that the latter locations are reasonably accurate.

Model	ρ
Global	.47
GT-Obj	.39
GT-Surr	.50
GT-Obj+Surr	**.54**
AL-Obj	.36
AL-Surr	.51
AL-Obj+Surr	**.54**

Table 3: Spearman correlations of the models with human semantic relatedness intuitions for the Pascal concepts.

For both the ground-truth (GT) and automated (AL) localization we observe an intriguing pattern: The models outperform Global when they rely on features extracted from the surroundings, whereas the object-only features perform relatively poorly. This shows how the distributional hypothesis transfers to images: The appearance of context is the most informative for semantics, similar to the case in text (in which it is the only information). This makes sense as semantically related objects share functionality, making them appear in similar contexts. The appearance of the object alone, on the other hand, has rather poor performance for semantics, even though semantically related objects can share characteristics, such as *wheels* or *fur*. The mix of object and surround into a Global representation only reduces performance. However, a concatenation of the two representations (Obj+Surr in the table) does improve results, confirming that semantically similar objects do share some similarity in their appearance, and that it is better not to mix properties of objects and properties of contexts.

To gain qualitative insights into what the models are doing, we looked at the similarity matrices obtained by comparing concepts with each other in terms of cosine similarity, shown in Figure 2. This enables checking how the overall model-based patterns compare to human intuitions. We focus on AL only because it has comparable performance to ground-truth segmentation, but it is obtained fully automatically, making it more useful for our purposes.

By looking first at the human similarity matrix, we notice the blocks corresponding to the three classes of *indoor objects, animals* and *vehicles*. While *animals* and *vehicles* are clearly discernible, the *indoor* class does not emerge with the same clarity, due to phenomena such as *chair* being also semantically related to *bus* and *boat* (buses and boats typically contain chairs), or *cat* being related to *sofa* (one of the most likely place where you can find one). More in general, the *indoor* class shares some semantic properties with both *animals* and *vehicles*, which makes its borders somewhat fuzzy.

Looking now at the visual similarity matrices, Global is capturing the same division into classes, albeit with some different patterns compared to humans: there is much more fuzziness in *animals* and *vehicles* than in the *indoor* class. AL-Obj is better at grouping together *animals* and *vehicles*, except for *bicycle* and *motorbike* which share many visual characteristics with *animals*. This is presumably because the spokes of the wheels contain many edges in different orientations just like animal fur, yielding similar visual words in SIFT space. For *indoor objects* AL-Obj finds little correlations. AL-Surr is capturing the three classes very clearly,

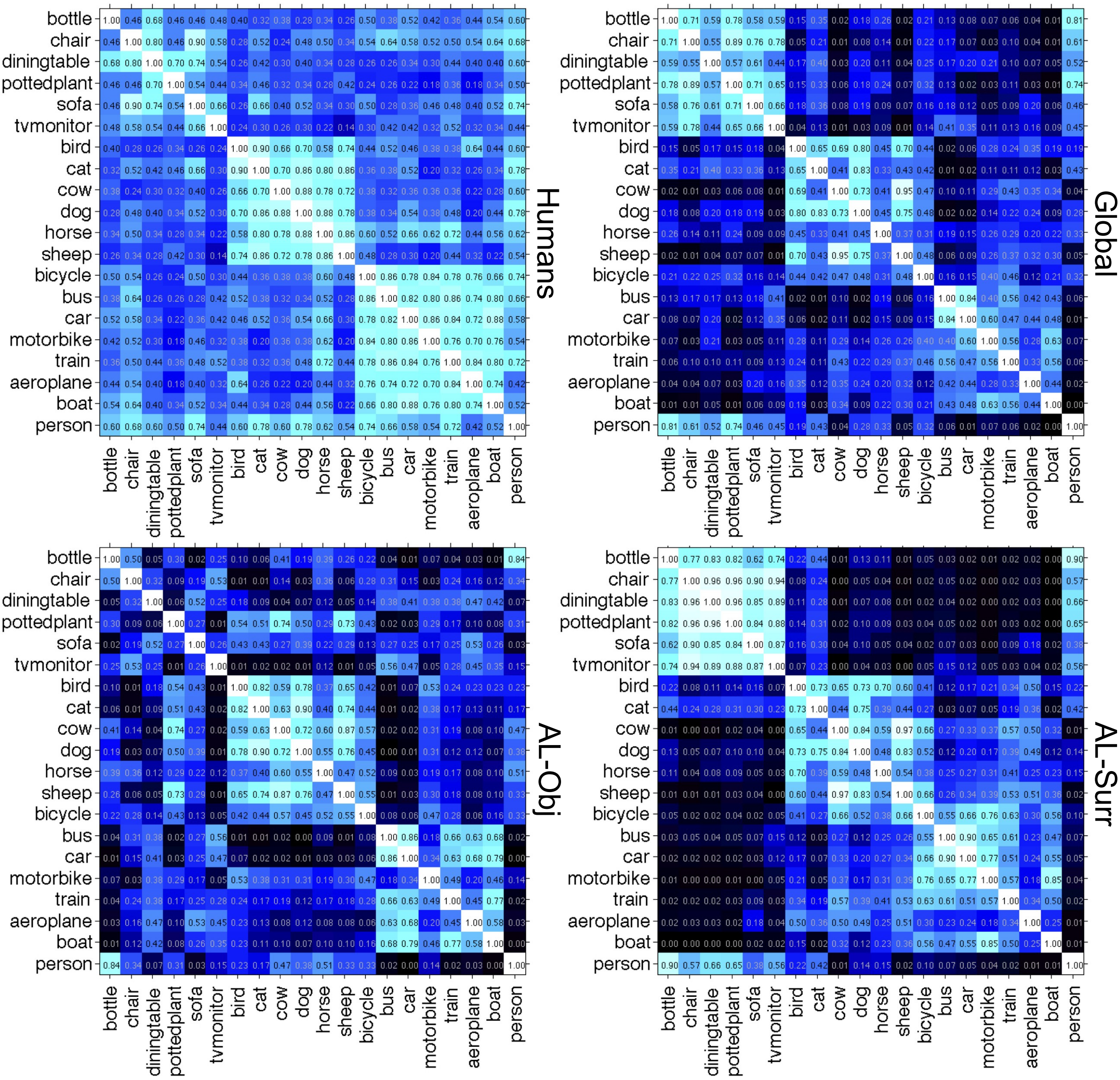

Figure 2: Similarity matrices for the human subjects (top left), Global (top right), AL-Obj (bottom left) and AL-Surr (bottom right). Lighter color cues higher similarity.

with just a bit more confusion between *animals* and *vehicles* (*aeroplane* being nearer to *animals* than to most *vehicles*). This model captures the *indoor* cluster particularly well, probably because "indoorness" is mostly a property of the surroundings of an object.

To conclude, we showed that the distributional hypothesis applied to visual features indeed holds: Semantically similar objects occur in similar contexts, just like in text. Additionally, the appearance of the object does contain extra information over the context alone but it is more helpful when the distinction between object and context is retained.

5. CONCLUSION

The novel idea of this work is to enrich text-based semantic analysis with features automatically extracted from images that are associated with words in tagged image collections. This goes against the current trend in multimedia and image analysis which advocates the use of text to improve image-related tasks by typically attempting to model the relation between specific images and specific textual passages. We introduced a fully fleshed-out model of image-based distributional vector extraction and integration. Our analysis confirmed that image-based distributional vectors contain sufficient information to capture a network of sensible word meaning relations, and that there are intriguing differences in the relations picked by the text- and image-based vectors, pointing to their complementarity. We showed that an integrated text-image semantic model significantly outperforms a state-of-the-art purely textual model. Moreover, we evaluated the impact that localizing an object in an image has on the image-based semantic representation of the meaning of the corresponding word.

Future work will focus, first of all, on improving the fusion of text- and image-based vectors, since the current results suggest that a good fusion strategy is crucial to attain significant improvements over text alone. A promising route to explore will be that of assigning different weights to the textual and visual components on a word-by-word basis, instead of doing it globally (for example, assigning more weights to the image-based vectors of more imageable words).

Exploiting localization gave very promising results in our last experiment. However, current localization algorithms are limited to objects for which segmented training data are available. Thus, we intend to explore possibly noisier unsupervised or weakly supervised localization methods, to scale up location-sensitive extraction of semantic vectors to represent thousands of word meanings. An interesting related question is whether localization will also help when harvesting meaning representations not only of objects but also attributes and events.

Last but not least, image-based and multimodal vectors should be evaluated in more semantic tasks and applications, investigating in particular whether the measures of semantic relatedness they provide are helpful for the analysis of text that is more strongly "grounded" in vision, such as captions, comments to videos and pictures or broadcast transcriptions.

6. ACKNOWLEDGMENTS

EB and MB were partially funded by a Google Research Award. JU and NS were partially funded by FP7 IP GLOCAL and FIRB S-PATTERNS projects. We thank Nam Khanh Tram for the help with the combination experiments.

7. REFERENCES

[1] M. Andrews, G. Vigliocco, and D. Vinson. Integrating experiential and distributional data to learn semantic representations. *Psych. Rev.*, 116(3):463–498, 2009.

[2] K. Barnard, P. Duygulu, D. Forsyth, N. de Freitas, D. Blei, and M. Jordan. Matching words and pictures. *J. Machine Learning Research*, 3:1107–1135, 2003.

[3] K. Barnard and M. Johnson. Word sense disambiguation with pictures. *Artificial Intelligence*, 167:13–30, 2005.

[4] M. Baroni, E. Barbu, B. Murphy, and M. Poesio. Strudel: A distributional semantic model based on properties and types. *Cognitive Science*, 34(2):222–254, 2010.

[5] M. Baroni and A. Lenci. Concepts and properties in word spaces. *Italian Journal of Linguistics*, 20(1):55–88, 2008.

[6] M. Baroni and A. Lenci. How we BLESSed distributional semantic evaluation. In *Proceedings of the EMNLP GEMS Workshop*, pages 1–10, 2011.

[7] T. Berg, A. Berg, and J. Shih. Automatic attribute discovery and characterization from noisy Web data. In *Proceedings of ECCV*, pages 663–676, 2010.

[8] S. Bergsma and R. Goebel. Using visual information to predict lexical preference. In *Proceedings of Recent Advances in Natural Language Processing*, pages 399–405, 2011.

[9] D. Blei, A. Ng, and M. Jordan. Latent Dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022, 2003.

[10] A. Bosch, A. Zisserman, and X. Munoz. Scene classification using a hybrid generative/discriminative approach. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 30(4):712–727, 2008.

[11] E. Bruni, G. Boleda, M. Baroni, and N. K. Tran. Distributional semantics in technicolor. In *Proceedings of the ACL 2012*, South Korea, 2012.

[12] E. Bruni, G. B. Tran, and M. Baroni. Distributional semantics from text and images. In *Proceedings of the EMNLP GEMS Workshop*, pages 22–32, Edinburgh, 2011.

[13] J. Caicedo, J. Ben-Abdallah, F. González, and O. Nasraoui. Multimodal representation, indexing, automated annotation and retrieval of image collections via non-negative matrix factorization. *Neurocomputing*, 76(1):50–60, 2012.

[14] G. Csurka, C. Dance, L. Fan, J. Willamowski, and C. Bray. Visual categorization with bags of keypoints. In *ECCV Workshop on Statistical Learning in Computer Vision*, pages 1–22, 2004.

[15] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *Proceedings of CVPR*, 2005.

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

[17] H. Escalante, C. Hérnadez, L. Sucar, and M. Montes. Late fusion of heterogeneous methods for multimedia image retrieval. In *ACM Conference on Multimedia Information Retrieval*, 2008.

[18] M. Everingham, L. van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The PASCAL Visual Object Classes (VOC) challenge. *International Journal of Computer Vision*, 2010.

[19] S. Evert. *The Statistics of Word Cooccurrences.* Dissertation, Stuttgart University, 2005.

[20] A. Farhadi, I. Endres, D. Hoiem, and D. Forsyth. Describing objects by their attributes. In *Proceedings of CVPR*, pages 1778–1785, 2009.

[21] A. Farhadi, M. Hejrati, M. A. Sadeghi, P. Young, C. Rashtchian, J. Hockenmaier, and D. Forsyth. Every picture tells a story: Generating sentences from images. In *Proceedings of ECCV*, 2010.

[22] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan. Object detection with discriminatively trained part based models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32:1627–1645, 2010.

[23] Y. Feng and M. Lapata. Visual information in semantic representation. In *Proceedings of HLT-NAACL*, pages 91–99, 2010.

[24] K. Grauman and T. Darrell. The pyramid match

kernel: Discriminative classification with sets of image features. In *Proceedings of ICCV*, pages 1458–1465, 2005.

[25] G. Kulkarni, V. Premraj, S. Dhar, S. Li, Y. Choi, A. C. Berg, and T. L. Berg. Baby talk: Understanding and generating simple image descriptions. In *Proceedings of CVPR*, 2011.

[26] C. H. Lampert, M. B. Blaschko, and T. Hofmann. Efficient Subwindow Search: A Branch and Bound Framework for Object Localization. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31:2129–2142, 2009.

[27] T. Landauer and S. Dumais. A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psych. Rev.*, 104(2):211–240, 1997.

[28] I. Laptev. Improving Object Detection with Boosted Histograms. *Image and Vision Computing*, 27:535–544, 2009.

[29] V. Lavrenko, R. Manmatha, and J. Jeon. A model for learning the semantics of pictures. In *NIPS*, 2003.

[30] S. Lazebnik, C. Schmid, and J. Ponce. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In *Proceedings of CVPR*, pages 2169–2178, 2006.

[31] C. W. Leong and R. Mihalcea. Going beyond text: A hybrid image-text approach for measuring word relatedness. In *Proceedings of IJCNLP*, pages 1403–1407, 2011.

[32] D. Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60(2), Nov. 2004.

[33] W. Lowe. Towards a theory of semantic space. In *Proceedings of CogSci*, pages 576–581, 2001.

[34] K. Lund and C. Burgess. Producing high-dimensional semantic spaces from lexical co-occurrence. *Behavior Research Methods*, 28:203–208, 1996.

[35] C. Manning, P. Raghavan, and H. Schütze. *Introduction to Information Retrieval.* Cambridge University Press, Cambridge, 2008.

[36] G. Miller and W. Charles. Contextual correlates of semantic similarity. *Language and Cognitive Processes*, 6(1):1–28, 1991.

[37] D. Moore and G. McCabe. *Introduction to the Practice of Statistics.* Freeman, New York, 5 edition, 2005.

[38] F. Moosmann, B. Triggs, and F. Jurie. Fast discriminative visual codebooks using randomized clustering forests. In *Neural Information Processing Systems*, pages 985–992, 2006.

[39] B. Murphy. *A Study of Notions of Participation and Discourse in Argument Structure Realisation.* Dissertation, Trinity College Dublin, 2007.

[40] D. Nister and H. Stewenius. Scalable recognition with a vocabulary tree. In *Proceedings of CVPR*, pages 2161–2168, 2006.

[41] S. Padó and M. Lapata. Dependency-based construction of semantic space models. *Computational Linguistics*, 33(2):161–199, 2007.

[42] D. Parikh and K. Grauman. Relative attributes. In *Proceedings of ICCV*, 2011.

[43] T.-T. Pham, N. Maillot, J.-H. Lim, and J.-P. Chevallet. Latent semantic fusion model for image retrieval and annotation. In *Proceedings of CIKM*, pages 439–443, 2007.

[44] R. Rapp. Word sense discovery based on sense descriptor dissimilarity. In *Proceedings of MT Summit*, pages 315–322, 2003.

[45] N. Rasiwasia, J. Costa Pereira, E. Coviello, G. Doyle, G. Lanckriet, R. Levy, and N. Vasconcelos. A new approach to cross-modal multimedia retrieval. In *ACM MM*, 2010.

[46] B. Riordan and M. Jones. Redundancy in perceptual and linguistic experience: Comparing feature-based and distributional models of semantic representation. *Topics in Cognitive Science*, 3(2):1–43, 2011.

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

[48] M. Sahlgren. An introduction to random indexing. `http://www.sics.se/~mange/papers/RI_intro.pdf`, 2005.

[49] M. Sahlgren. *The Word-Space Model.* Dissertation, Stockholm University, 2006.

[50] H. Schütze. *Ambiguity Resolution in Natural Language Learning.* CSLI, Stanford, CA, 1997.

[51] J. Sivic and A. Zisserman. Video Google: A Text Retrieval Approach to Object Matching in Videos. In *Proceedings of ICCV*, 2003.

[52] C. G. M. Snoek. *The Authoring Metaphor to Machine Understanding of Multimedia.* PhD thesis, University of Amsterdam, 2005.

[53] P. Turney and P. Pantel. From frequency to meaning: Vector space models of semantics. *Journal of Artificial Intelligence Research*, 37:141–188, 2010.

[54] J. R. R. Uijlings, A. W. M. Smeulders, and R. J. H. Scha. Real-time Visual Concept Classification. *IEEE Transactions on Multimedia*, 12, 2010.

[55] K. E. A. van de Sande, T. Gevers, and C. G. M. Snoek. Evaluating Color Descriptors for Object and Scene Recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2010.

[56] K. E. A. van de Sande, J. Uijlings, T. Gevers, and A. Smeulders. Segmentation as Selective Search for Object Recognition. In *Proceedings of ICCV*, 2011.

[57] P. Viola and M. J. Jones. Robust real-time face detection. *International Journal of Computer Vision*, 57:137–154, 2004.

[58] D. Vreeswijk, B. Huurnink, and A. Smeulders. Text and image subject classifiers: dense works better. In *ACM MM*, 2011.

[59] G. Wang and D. Forsyth. Joint learning of visual attributes, object classes and visual saliency. In *Proceedings of ICCV*, 2009.

[60] J. Yang, Y.-G. Jiang, A. G. Hauptmann, and C.-W. Ngo. Evaluating bag-of-visual-words representations in scene classification. In *Multimedia Information Retrieval*, pages 197–206, 2007.

[61] Z.-J. Zha, L. Yang, T. Mei, M. Wang, and Z. Wang. Visual query suggestion. In *Proceedings of ACM Multimedia*, 2009.