Combining Text Semantics and Image Geometry to Improve Scene Interpretation

Dennis Medved[1], Fangyuan Jiang[2], Peter Exner[1], Magnus Oskarsson[2], Pierre Nugues[1] and Kalle Åström[2]
[1]Department of Computer Science, Lund University, Lund, Sweden
[2]Department of Mathematics, Lund University, Lund, Sweden

Keywords: Semantic Parsing, Relation Extraction from Images, Machine Learning.

Abstract: In this paper, we describe a novel system that identifies relations between the objects extracted from an image. We started from the idea that in addition to the geometric and visual properties of the image objects, we could exploit lexical and semantic information from the text accompanying the image. As experimental set up, we gathered a corpus of images from Wikipedia as well as their associated articles. We extracted two types of objects: human beings and horses and we considered three relations that could hold between them: *Ride*, *Lead*, or *None*. We used geometric features as a baseline to identify the relations between the entities and we describe the improvements brought by the addition of bag-of-word features and predicate–argument structures we derived from the text. The best semantic model resulted in a relative error reduction of more than 18% over the baseline.

1 INTRODUCTION

A large percentage of queries to retrieve images relate to people and objects (Markkula and Sormunen, 2000; Westman and Oittinen, 2006) as well as relations between them: the 'story' within the image (Jörgensen, 1998). Although the automatic recognition, detection and segmentation of objects in images has reached remarkable levels of accuracy, reflected by the Pascal VOC Challenge evaluation (Carreira and Sminchisescu, 2010; Felzenszwalb et al., 2010; Ladicky et al., 2010), the identification of relations is still a territory that is yet largely unexplored. Notable exceptions include Chen et al. (2012) and Myeong et al. (2012). The identification of these relations, though, would enable users to search images illustrating two or more objects more accurately.

Relations between objects within images are often ambiguous and captions are intended to help us in their interpretation. As human beings, we often have to read the caption or the surrounding text to understand what happened and the nature of the relations between the entities. This combined use of text and images has been explored in automatic interpretation mostly in the form of bag of words, see Sect. 2. This approach might be inadequate however, as bags of words do not take the word or sentence context into account. This model inadequacy formed the starting idea of this project: As we focused on relations in images, we tried to model their counterparts in the text and reflect them not only with bags of words but also in the form of predicate–argument structures.

2 RELATED WORK

To the best of our knowledge, no work has been done to identify relations in images using a combined analysis of image and text data. There are related works however:

Paek et al. (1999) combined image segmentation with a text-based classifier using image captions as input. They used bags of words and applied a $TF \cdot IDF$ weighting on the text. The goal was to label the images as either taken indoor or outdoor. They improved the results by using both text and image information together, compared to using only one of the classifiers.

Deschacht and Moens (2007) used a set of 100 image-text pairs from *Yahoo! News* and automatically annotated the images utilizing the associated text. The goal was to detect the presence of specific humans, but also more general objects. They analyzed the im-

Medved D., Jiang F., Exner P., Oskarsson M., Nugues P. and Aström K..
Combining Text Semantics and Image Geometry to Improve Scene Interpretation.
DOI: 10.5220/0004752004790486
In *Proceedings of the 3rd International Conference on Pattern Recognition Applications and Methods* (ICPRAM-2014), pages 479-486
ISBN: 978-989-758-018-5

age captions to find named entities. They also derived information from discourse segmentation, which was used to determine the saliency of entities.

Moscato et al. (2009) used a large corpus of French news articles, composed of a text, images, and image captions. They combined an image detector to recognize human faces and logos, with a named entity detection in the text. The goal was to correctly annotate the faces and logos found in the images. The images were not annotated by humans, instead named entities in the captions were used as the ground truth, and the classification was based on the articles.

Marszalek and Schmid (2007) used a semantic network and image labels to integrate prior knowledge of inter-class relationships in the learning step of a classifier to achieve better classification results. All of these works combined text and image analysis for classification purposes, but they did not identify relations in the images. Another area of related work is the generation of natural language descriptions of an image scene, see Gupta et al. (2012) and Kulkarni et al. (2011).

3 DATA SET AND EXPERIMENTAL SETUP

The internet provides plenty of combined sources of images and text including news articles, blogs, and social media. Wikipedia is one of such sources that, in addition to a large number of articles, is a substantial repository of images illustrating the articles. As of today, the English version has over 4 million articles and about 2 million images (Wikipedia, 2012). It is not unusual for editors to use an image for more than one article, and an image can therefore have more than one article or caption associated with it.

We gathered a subset of images and articles from Wikipedia restricted to two object categories: *Horse* and *Human*. We extracted the articles containing the keywords *Horse* or *Pony* and we selected their associated images. This resulted into 901 images, where 788 could be used. Some images were duplicates and some did not have a valid article associated with them.

An image connected to the articles with the words *Horse* or *Pony* does not need to contain a real horse. It can depict something associated with the words for example: a car, a statue, or a painting. Some of the images also include humans, either interacting with the horse or just being part of the background, see Figure 1 for examples. An image can therefore have none or multiple horses, and none or multiple humans.

We manually annotated the horses and humans in the images with a set of possible relations: *Ride*, *Lead*, and *None*. *Ride* and *Lead* are when a human is riding or leading a horse and *None* is an action that is not *Ride* or *Lead* including no action at all. The annotation gave us the number of respective humans and horses, their sizes and their locations in the image.

We processed the articles with a semantic parser (Exner and Nugues, 2012), where the output for each word is its lemma and part of speech, and for each sentence, the dependency graph and predicate-argument structures it contains. We finally applied a coreference solver to each article.

4 VISUAL PARSING

As our focus was to investigate to what extent the use of combinations of text and visual cues could improve the interpretation or categorization precision, we set aside the automatic detection of objects in the images. We manually identified the objects within the images by creating bounding boxes around horses and humans. We then labeled the interaction between the human-horse pair if the interaction corresponded to *Lead* or *Ride*. The *None* relationships were left implicit. It resulted in 2,235 possible human-horse pairs in the images, but the distribution of relations was quite heavily skewed towards the None relation. The Lead relation had significantly fewer examples; see Table 1.

Table 1: The number of different objects in the source material.

Item	Count
Extracted images	901
Usable images	788
Human-horse pairs	2,235
Relation: *None*	1,935
Relation: *Ride*	233
Relation: *Lead*	67

The generation of the bounding boxes could be produced automatically by an object detection algorithm trained on the relevant categories (in our case people and horses) such as e.g. the deformable part-based model described in Felzenszwalb et al. (2010). This would have enabled us to skip the manual detection step, but as our focus in this paper lies elsewhere we opted not to do this.

5 SEMANTIC PARSING

We used the Athena parsing framework (Exner and Nugues, 2012) in conjunction with a coreference

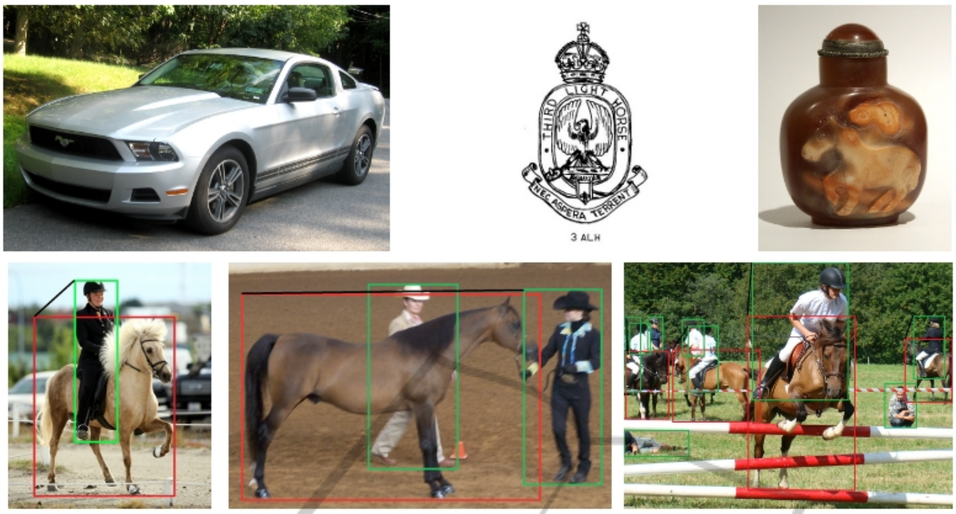
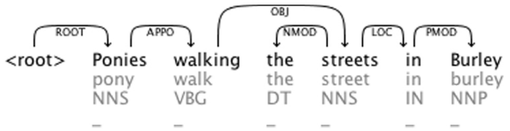


Figure 1: The upper row shows: A Ford Mustang, the 3rd Light Horse Regiment hat badge, and a snuff bottle. The lower row shows: A human riding a horse, one human leading the horse and one bystander, and seven riders and two bystanders. Bounding boxes are displayed.

solver (Stamborg et al., 2012) to parse the Wikipedia articles. For each word, the parser outputs its lemma and part of speech (POS). In addition, the parser produces a dependency graph with labeled edges for each sentence as well as the predicates it contains and their arguments. For each article, we also identify the words or phrases that refer to a same entity i.e. words or phrases that are coreferent.

Figure 2 shows the dependency graph and the predicate–argument structure of the caption: *Ponies walking the streets in Burley*[1].



	Ponies	walking	the	streets	in	Burley
walk.01	A0		A1			

Figure 2: A representation of a parsed sentence: The upper part shows the syntactic dependency graph and the lower part shows the predicate, *walk*, and its two arguments the parser has extracted: *Ponies* and *the streets in Burley*.

5.1 Predicates

The semantic parser uses the PropBank (Palmer et al., 2005) nomenclature, where the predicate sense is explicitly shown as a number added after the word. The sentence in Figure 2 contains one predicate: walk.01 with its two arguments A0 and A1, where A0 corresponds to the walker and A1, the path walked.

PropBank predicates can also have modifying arguments denoted with the prefix "AM-". There exist 14 different types of modifiers in PropBank such as:

AM-DIR: shows motion along some path,

AM-LOC: indicates where the action took place, and

AM-TMP: shows when the action took place.

5.2 Coreferences

We applied a coreference resolution to create sets of coreferring mentions as with *the rider* and the two *he* in this caption:

> If *the rider* has a refusal at the direct route *he* may jump the other B element without additional penalty than what *he* incurred for the refusal.[2]

The phrase *the rider* is the first mention of an entity in the coreference chain. It usually contains most information in the chain. We use it together with POS information and we substitute coreferent words with this mention in a document, although this is mostly useful with pronouns. The modified documents can thereafter be used with different lexical features.

[1] http://en.wikipedia.org/wiki/New_Forest, retrieved November 9, 2012.

[2] http://en.wikipedia.org/wiki/Eventing, retrieved November 9, 2012.

6 FEATURE EXTRACTION

We used classifiers with visual and semantic features to identify the relations. The visual features formed a baseline system. We then added semantic features to investigate this improvement over the baseline.

6.1 Visual Features

The visual parsing annotation provided us with a set of objects within the images and their bounding boxes defined by the coordinates of the center of each box, its width, and height.

To implement the baseline, we derived a larger set of visual features from the bounding boxes, such as the overlapping area, the relative positions, etc, and combinations of them. We ran an automatic generation of feature combinations and we applied a feature selection process to derive our visual feature set. We evaluated the results using cross-validation. However, as the possible number of combinations was very large, we had to discard manually a large part of them. Once stabilized, the baseline feature set remained unchanged while developing and testing lexical features. It contains the following features:

F_Overlap Boolean feature describing whether the two bounding boxes overlap or not.

F_Distance numerical feature containing the normalized length between the centers of the bounding boxes.

F_Direction(8) nominal feature containing the direction of the human relative the horse, discretized into eight directions.

F_Angle numerical feature containing the angle between the centers of the boxes.

F_OverlapArea numerical feature containing the size of the overlapping area of the boxes.

F_MinDistanceSide numerical feature containing the minimum distance between the sides of the boxes.

F_AreaDifference numerical feature containing the quotient of the areas.

We used logistic regression and to cope with non-linearities, we used pairs of features to emulate a quadratic function. The three following features are pairs involving a numerical and a Boolean features, creating a numerical feature. The Boolean feature is used as a step function: if it is false, the output is a constant; if it is true, the output is the value of the numeric feature.

F_Distance+F_LowAngle(7) numerical feature, F_LowAngle is true if the difference in angle is less than 7°.

F_Angle+F_LowAngle(7) numerical feature.

F_Angle+F_BelowDistance(100) numerical feature, F_BelowDistance(100) is true if the distance is less than 100.

Without these feature pairs, the classifier could not correctly identify the *Lead* relation and the F_1 value for it was 0. With these features, F_1 increased to 0.29. Table 2 shows the recall, precision, and F_1 for the three relations using visual features. Table 3, shows the corresponding confusion matrix.

Table 2: Precision, recall and F_1 for visual features.

	Precision	Recall	F_1
None	0.9472	0.9648	0.9559
Ride	0.7685	0.7553	0.7619
Lead	0.4285	0.2239	0.2941
Mean			0.6706

Table 3: The confusion matrix for visual features.

		Predicted class		
		None	*Ride*	*Lead*
Actual class	*None*	1867	49	19
	Ride	56	176	1
	Lead	48	4	15

6.2 Semantic Features

We extracted the semantic features from the Wikipedia articles. We implemented a selector to choose the size of the input between: complete articles, partial articles (the paragraph that is the closest to an image), captions, and file names. The most specific information pertaining to an image is found in the caption and the file name, followed by the partial article, and finally, the whole article.

6.2.1 Bag-of-Words Features

A bag-of-word (BoW) feature was created for each of the four different inputs. A BoW feature is represented by a vector of weighted word frequencies. The different versions have separate settings and dictionaries. We also used a combined bag-of-word feature vector consisting of the concatenation of the partial article, caption, and filename feature vectors.

The features have a filter that can exclude words that are either too common, or not common enough, based on their frequency, controlled by a threshold. We used a $TF \cdot IDF$ weighting on the included words.

We used file names as one of the inputs, as it is common to have a long descriptive names of the images in Wikipedia. However, they are not as standardized as the captions. Some images have very long descriptive titles; others were less informative, for example: "DMZ1.jpg". The file names were not semantically parsed, but we defined a heuristic algorithm, which was used to break down the file name strings into individual words.

6.2.2 Predicate Features

Instead of using all of the words in a document, we used information derived from the predicate–argument structure to filter out more relevant terms. We created a feature that only used the predicate names and their arguments as input. The words that are not predicates, or arguments to the predicates, are removed as input to the feature. The arguments can be filtered depending on their type, for example A0, A1, or AM-TMP. We can either consider all of the words of the arguments, or only the heads.

As for the BoW, we created predicate features with articles, partial articles, and captions as input. We never used the file names, because we could not carry a semantic analysis on them. We also created a version of the predicate-based features that we could filter further on the basis of a list of predicate names, including only predicates present in a predefined list, specified by regular expressions.

7 CLASSIFICATION

To classify the relations, we used the LIBLINEAR (Fan et al., 2008) package and the output probabilities over all the classes. The easiest way to classify a horse-human pair is to take the corresponding probability vector and pick the class with the highest probability. But sometimes the probabilities are quite equal and there is no clear class to chose. We selected a threshold using cross-validation. If the maximum probability in the vector is not higher than the threshold, the pair is classified as *None*. We observed that because *None* represents a collection of actions and nonaction, it is more likely to be the true class when *Ride* and *Lead* have low probabilities.

Even with the threshold, this scheme can classify two or more humans as riding or leading the same horse. Although possible, it is more likely that only one person is riding or leading the horse at a time. Therefore we added constraints to the classification: a horse can only have zero or one rider, and zero or one leader. For each class, only the most probable human is chosen, and only if it is higher than the threshold.

For each human-horse pair, the predicted class is compared to the actual class. The information derived from this can be used to calculate the precision, recall, and F_1 for each class. The arithmetic mean of the three F_1 values is calculated, and can be used as a comparison value. We also computed the number of correct classifications and a confusion matrix.

8 SYSTEM ARCHITECTURE

Figure 3 summarizes the architecture of the whole system:

1. Wikipedia is the source of the images and the articles. The text annotation uses the Wiki markup language.
2. Image analysis: placement of bounding boxes, classification of objects and actions. This was done manually, but could be replaced by an automatic system.
3. Text selector between: the whole articles, paragraphs that are the closest to the images, filenames, or captions.
4. Semantic parsing of the text, see Section 5.
5. Extraction of feature vectors based on the bounding boxes and the semantic information.
6. Model training using logistic regression from the LIBLINEAR package. This enables us to predict probabilities for the different relations.
7. Relation classification using probabilities and constraints.

9 RESULTS

We used the L2-regularized logistic regression (primal) solver from the LIBLINEAR package and we evaluated the results of the classification with the different feature sets starting from the baseline geometric features and adding lexical features of increasing complexity. We carried out a 5-fold crossvalidation.

We evaluated permutations of features and settings and we report the set of combined BoW features that yielded the best result. Table 4 shows an overview of the results:

- The baseline corresponds to the geometrical features; we obtained a mean F_1 of 0.67 with them;

Table 4: An overview of the results, with their mean F_1-value, difference and relative error reduction from the baseline mean F_1-value.

		Mean of F_1	Difference (pp)	Relative error reduction (%)
Baseline		0.6706	0.00	0.00
BoW	Articles	0.6779	0.73	2.22
	Partial articles	0.6818	1.12	3.40
	Captions	0.6829	1.23	3.73
	Filenames	0.6802	0.96	2.91
	Combination	0.7132	4.26	12.9
Predicate	Articles	**0.7318**	**6.12**	**18.6**
	Partial articles	0.6933	2.27	6.89
	Captions	0.6791	0.85	2.58
	Articles + Words	0.6830	1.24	3.76
	Articles + Coref	0.7280	5.74	17.4

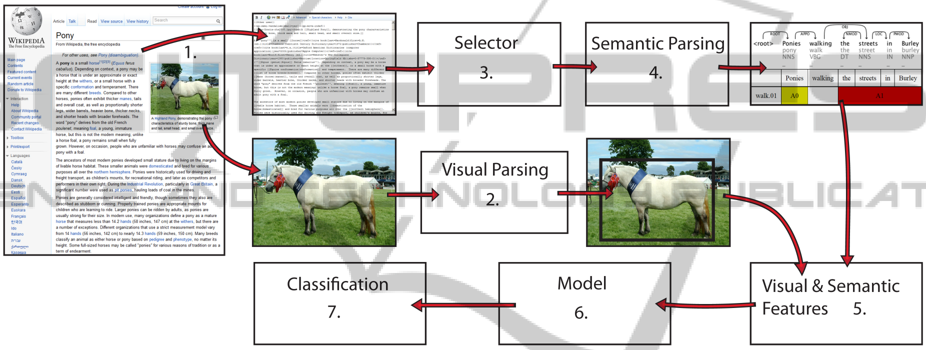


Figure 3: An overview of the system design, see Section 8 for description.

- BoW corresponds to the baseline features and the bag-of-word features described in Sect. 6.2.1; whatever the type of text we used as input, we observed an improvement. We obtained the best results with a concatenation of the partial article, caption, and filename (combination, $F_1 = 0.71$);
- predicate corresponds to the baseline features and the predicate feature vector described in Sect. 6.2.2. Predicate features using only one lexical feature vector from the article text gave better results than combining different portions of the text ($F_1 = 0.73$).

Our best feature set is the predicate features utilizing whole articles as input. It achieves a relative error reduction of 18.6 percent compared to baseline.

Tables 2 and 3 show the detailed results of the baseline with the geometric features only. Tables 5 and 6 show the results of the best BoW feature combination: a concatenation of the feature vectors from the inputs: partial articles, captions, and filenames. Tables 7 and 8 show the result of the best predicate features.

10 DISCUSSION

Classifying the *Lead* relation with geometric features with only bounding boxes as the input revealed quite difficult. There is indeed very little visual difference between standing next to a horse and leading it. We were not able to classify any *Lead* correctly until we added the combination features.

For single BoW features, the captions gave the best result, followed by partial articles, filenames, and lastly articles. The order of the results was what we expected, based on how specific information the features had about the images. But for the predicate features, the order was reversed: articles produced the best result, followed by partial articles, and captions.

Using a specific list of predicates did not produce good results although, depending on the list, results vary greatly. Using a list with the words: *ride*, *lead*, *pull*, and *race*, with articles as input, gave the best result, but Table 4 shows a relative drop of 4.88 compared to no filtering. The negative results could possibly be explained by the fact that it is not common to explicitly describe the relations in the images, and only utilizing keywords such as *ride* is of little help.

Applying coreference resolution on the documents lowered the results. Table 4 shows a relative drop of 0.38 if applied on the predicate feature based on articles. Despite these negative results, we still believe that solving coreferences could improve the results. The solver was designed to be used with another set of semantic information. To be able to use the solver, we altered its source code and possibly made it less accurate. We checked manually coreference chains and we could observe a significant number of faulty examples, leading us to believe that the output quality of the solver left to be desired.

Table 5: Precision, recall, and F_1 for the concatenation of BoW features with the inputs: partial articles, captions and filenames.

	Precision	Recall	F_1
None	0.9638	0.9638	0.9638
Ride	0.7642	0.8626	0.8104
Lead	0.5135	0.2835	0.3653
Mean			0.7132

Table 6: The confusion matrix for BoW for the concatenation of BoW features with the inputs: partial articles, captions and filenames.

		Predicted class		
		None	*Ride*	*Lead*
Actual class	*None*	1865	57	13
	Ride	27	201	5
	Lead	43	5	19

Table 7: Precision, recall and F_1 for predicate feature on articles.

	Precision	Recall	F_1
None	0.9745	0.9498	0.9620
Ride	0.7301	0.9055	0.8084
Lead	0.4500	0.4029	0.4251
Mean			0.7318

Table 8: The confusion matrix for predicate feature on articles.

		Predicted class		
		None	*Ride*	*Lead*
Actual class	*None*	1838	70	27
	Ride	16	211	6
	Lead	32	8	27

11 CONCLUSIONS AND FUTURE WORK

We designed a supervised classifier to identify relations between pairs of objects in an image. As input to the classifier, we used geometric, bag-of-words, and semantic features. The results we obtained show that semantic information, in combination with geometric features, proved useful to improve the classification of relations in the images. Table 4 shows that the relative error reduction is 12.9 percent by utilizing a combination of bag-of-words features. An even greater improvement is made using predicate information with an relative error reduction of 18.6 percent compared to baseline.

Coreference resolution lowered the performance, but the interface between the semantic parser and the coreference solver was less than optimal. There is room for improvement regarding this solver, either with the interface to the semantic parser or with to another solver. It could also be interesting to try other types of classifiers, not just logistic regression, and see how they perform.

Using automatically annotated images as input to the program could be relatively easily implemented and would automate all the steps in the system. A natural continuation of the work is to expand the number of objects and relations. Felzenszwalb et al. (2010), for example, use 20 different classifiers for common objects: cars, bottles, birds, etc. All, or a subset of it, could be chosen as the objects, together with some common predicates between the objects as the relations.

It would also be interesting to try other sources of images and text than Wikipedia: either using other resources available online or creating a new database with images captioned with text descriptions. Another interesting expansion of the work would be to map entities found in the text with objects found in the image. For example, if a caption includes the name of a person, one could create a link between the image and information about the entity.

ACKNOWLEDGEMENTS

This research was supported by Vetenskapsrådet, the Swedish research council, under grant 621-2010-4800 and the Swedish e-science program: eSSENCE.

REFERENCES

Carreira, J. and Sminchisescu, C. (2010). Constrained Parametric Min-Cuts for Automatic Object Segmentation. In *IEEE International Conference on Computer Vision and Pattern Recognition*.

Chen, N., Zhou, Q.-Y., and Prasanna, V. (2012). Understanding web images by object relation network. In *Proceedings of the 21st international conference on World Wide Web*, WWW '12, pages 291–300, New York, NY, USA. ACM.

Deschacht, K. and Moens, M.-F. (2007). Text analysis for automatic image annotation. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 1000–1007, Prague.

Exner, P. and Nugues, P. (2012). Constructing large proposition databases. In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul.

Fan, R.-E., Chang, K.-W., Hsieh, C.-J., Wang, X.-R., and Lin, C.-J. (2008). LIBLINEAR: A library for large linear classification. *Journal of Machine Learning Research*, 9:1871–1874.

Felzenszwalb, P. F., Girshick, R. B., McAllester, D., and Ramanan, D. (2010). Object detection with discriminatively trained part based models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(9):1627–1645.

Gupta, A., Verma, Y., and Jawahar, C. (2012). Choosing linguistics over vision to describe images. In *Proc. of the twenty-sixth AAAI conference on artificial intelligence*.

Jörgensen, C. (1998). Attributes of images in describing tasks. *Information Processing and Management*, 34(2-3):161–174.

Kulkarni, G., Premraj, V., Dhar, S., Siming, L., Choi, Y., Berg, A., and Berg, T. (2011). Baby talk: Understanding and generating image descriptions. In *Proc. Conf. Computer Vision and Pattern Recognition*.

Ladicky, L., Russell, C., Kohli, P., and Torr, P. H. S. (2010). Graph cut based inference with co-occurrence statistics. In *Proceedings of the 11th European conference on Computer vision: Part V*, ECCV'10, pages 239–253, Berlin, Heidelberg. Springer-Verlag.

Markkula, M. and Sormunen, E. (2000). End-user searching challenges indexing practices in the digital newspaper photo archive. *Information retrieval*, 1(4):259–285.

Marszalek, M. and Schmid, C. (2007). Semantic hierarchies for visual object recognition. In *Proc. Conf. Computer Vision and Pattern Recognition*.

Moscato, V., Picariello, A., Persia, F., and Penta, A. (2009). A system for automatic image categorization. In *Semantic Computing, 2009. ICSC'09. IEEE International Conference on*, pages 624–629. IEEE.

Myeong, H., Chang, J. Y., and Lee, K. M. (2012). Learning object relationships via graph-based context model. In *CVPR*, pages 2727–2734.

Paek, S., Sable, C., Hatzivassiloglou, V., Jaimes, A., Schiffman, B., Chang, S., and Mckeown, K. (1999). Integration of visual and text-based approaches for the content labeling and classification of photographs. In *ACM SIGIR*, volume 99.

Palmer, M., Gildea, D., and Kingsbury, P. (2005). The Proposition Bank: an annotated corpus of semantic roles. *Computational Linguistics*, 31(1):71–105.

Stamborg, M., Medved, D., Exner, P., and Nugues, P. (2012). Using syntactic dependencies to solve coreferences. In *Joint Conference on EMNLP and CoNLL - Shared Task*, pages 64–70, Jeju Island, Korea. Association for Computational Linguistics.

Westman, S. and Oittinen, P. (2006). Image retrieval by end-users and intermediaries in a journalistic work context. In *Proceedings of the 1st international conference on Information interaction in context*, pages 102–110. ACM.

Wikipedia (2012). Wikipedia statistics English. http://stats.wikimedia.org/EN/TablesWikipediaEN.htm.