Learning Visual Compound Models from Parallel Image-Text Datasets

Jan Moringen[1], Sven Wachsmuth[1], Sven Dickinson[2], and Suzanne Stevenson[2]

[1] Bielefeld University
{jmoringe,swachsmu}@techfak.uni-bielefeld.de
[2] University of Toronto
{sven,suzanne}@cs.toronto.edu

Abstract. In this paper, we propose a new approach to learn structured visual compound models from shape-based feature descriptions. We use captioned text in order to drive the process of grouping boundary fragments detected in an image. In the learning framework, we transfer several techniques from computational linguistics to the visual domain and build on previous work in image annotation. A statistical translation model is used in order to establish links between caption words and image elements. Then, compounds are iteratively built up by using a mutual information measure. Relations between compound elements are automatically extracted and increase the discriminability of the visual models. We show results on different synthetic and realistic datasets in order to validate our approach.

1 Introduction

Parallel datasets are an interesting source for learning models that offer semantic access to huge collections of unstructured data. Given a corpus of paired images and captions, we aim to learn structured visual appearance models that can be used to automatically annotate images newly observed among other applications. In related work, a number of approaches already have been reported. These can be differentiated by the kind of image representation used, the kind of labelling expected, and the word-image association model applied. Barnard et al. [1,2] start from a blob representation of the image which is provided by a general image segmentation algorithm, and construct a vocabulary of visual words from the color and texture features extracted for each blob. They have explored different modelling approaches. First, they apply a generative hierarchical co-occurrence model proposed by Hofmann [3] that treats blobs and words as conditionally independent given a common topic. Thus, the visual correspondent of a word is not explicitly localized in an image. A second approach applies a statistical translation model (*Model 1*) of Brown et al. [4] which includes a set of alignment variables that directly link words to single blobs in the image. A third approach, followed by Blei & Jordan, combines both approaches using *latent dirichlet allocation* (LDA) [1]. Our approach strongly builds on the translation model approach but extends it towards compounds of shaped-based visual features.

G. Rigoll (Ed.): DAGM 2008, LNCS 5096, pp. 486–496, 2008.

Further work by Berg et al. [5] finds correspondences of image faces and caption names, but needs an initial set of unique assignments in order to drive the model construction process. Caneiro et al. [6] start from bags of localized image features and learn a Gaussian mixture model (GMM) for each semantic class which is provided by a weak labelling. They do not need any previous segmentation, but rely on the assumption that each caption word relates to the image. In contrast to the structureless bag-of-features model, Crandall & Huttenlocher [7] refer to the problem of learning part-based spatial object models from weakly labelled datasets. However, they assume a single label for each image that refers to the depicted compound object. Opelt et al. [8] propose a shape-based compound model that builds on class-discriminative boundary fragments. Each boundary fragment is defined relative to the object centroid. This provides a straightforward method for dealing with the object's spatial layout, but implies further constraints on training data because, besides the label, a bounding box needs to be specified for each image instance.

The work described in this paper is more tightly related to Jamieson et al. [9]. There, visual compound models are represented as collections of localized SIFT features. A visual vocabulary is learned on a separate data set. Initial feature-word correspondences are established by using a translation model similar to Barnard et al. [1]. Compound models are generated in an iterative search process that optimizes the translation model. Jamieson et al. [10] further improve the compound model by adding spatial relations of localized SIFT features, using pairs of features for initialization, and applying a more efficient correspondence model rather than a full-blown translation model. As impressive as these results are, the approach circumvents certain problems occurring for more general object categories. Many categories need to be characterized by their global shape rather than local gradient statistics. This includes the treatment of segmentation issues.

In the following, we will describe our approach to learning structured shape-models from image-caption pairs. It builds on some ideas already discussed by Wachsmuth et al. [11]. Again, initial correspondences are established by applying a statistical translation model. For learning visual compound models the framework of Melamed [12] is applied that is able to re-combine recurring sub-parts. Several new ideas are introduced in order to apply the general framework to shape-based representations. Boundary fragments are used for generating a visual vocabulary, the translation model is used in order to optimize detection thresholds of visual words, and compound models are extended by adding spatial relations between boundary fragments. The approach is evaluated on two datasets of image-caption pairs with different characteristics and shows its applicability for annotation tasks.

2 Image Representation

As mentioned before, we represent images by a set of localized visual words which are learned over shape-based features extracted from training images. Following Opelt et al. [8], we extract so called *boundary fragments* that are

connected components of region boundaries. In order to gain a more global measure of boundary pixels, we overlay several region segmentations using the EDISON system described by [13]. The accumulated boundary pixels are thresholded to create edge images, from which boundary fragments are extracted by concatenating edge pixels starting from randomly chosen seed pixels. However, in order to only select highly discriminative features, fragments are rejected if they (i) consist of too few pixels, (ii) consist of too many pixels, (iii) are not sufficiently curved.

Boundary fragments lend themselves well to fault-tolerant shape recognition using chamfer matching as described by Borgefors [14]. In our approach, we exploit this property for two purposes: (i) using chamfer matching, we locate previously built features in unknown images as described in [14]; and (ii) we apply chamfer matching as a distance metric between boundary fragments:

$$d_{\text{symm}}(f_1, f_2) := d_{\text{frag}}(f_1, f_2) + d_{\text{frag}}(f_2, f_1) \quad (1)$$
$$\text{where } d_{\text{frag}}(f_1, f_2) := \min_{T \in S} d_{\text{edge}}(f_1, T, I_{f_2}).$$

S is a discrete set of transformations[1] that is applied to boundary fragment f_1 when overlaying it over an image during chamfer matching, I_{f_2} is a bitmap-representation of the fragment f_2, and d_{edge} is the edge distance as described in [14].

In order to build up the visual vocabulary, we use a clustering approach on boundary fragments.

Definition 1. *A **cluster** is a triple (F, r, h) consisting of a set $F = \{f_1, \ldots, f_n\}$ of boundary fragments, a representative boundary fragment $r \in F$ with minimal distance d_{symm} to all other fragments in F, and a threshold $h \in \mathbb{R}_+$.*

Following Opelt et al. [8] we apply an agglomerative clustering technique, because the fact that boundary fragments do not form a vector space rules out methods such as k-means. The distance function[2] between two clusters l_1, l_2 can be derived from the distance between boundary fragments (Eq. 1) by

$$d_{\max}(l_1, l_2) := \max_{f_1 \in F_1, f_2 \in F_2} d_{\text{symm}}(f_1, f_2) \text{ where } l_k = (F_k, r_k, h_k), k \in \{1, 2\} \quad (2)$$

As it requires a large number of expensive computations of the edge distance, $d_{\max}$ cannot be used to cluster huge amounts of boundary fragments. To remedy this problem, we take the following two-step approach:

1. boundary fragments are agglomeratively clustered into classes of roughly the same overall shape using a computationally cheap distance function that compares variances along principal directions.
2. these classes of boundary fragments are further refined according to the distance function $d_{\max}$ by using agglomerative clustering on each subset.

[1] Here we only allow translations.

[2] Actually, a whole family of distance functions can be obtained by using aggregation functions like minimum or average, instead of maximum.

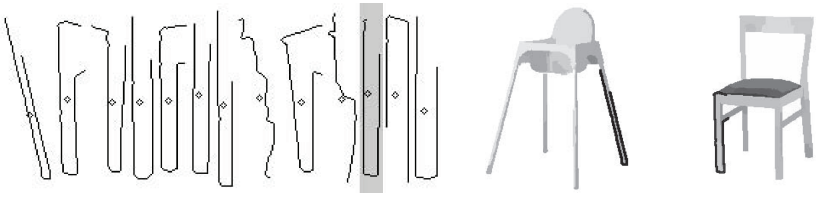


Fig. 1. Left: A cluster of boundary fragments. The highlighted boundary fragment represents the cluster. Right: Training images from which the two leftmost boundary fragments of the cluster have been extracted.

Besides building classes of similar fragments (see Fig. 1 for an example) for the translation model, each cluster also provides a detector for its corresponding class in unseen images. The detector exploits the representative boundary fragment as well as the threshold value h and is explained in more detail in Sec. 5.

3 Translation Model

We use a simple statistical machine translation model of the sort presented by Brown et al. [4] as *Model 1*. This model exploits co-occurences of visual words $\mathbf{c} = \{c_1, \ldots, c_M\}$ and image caption words $\mathbf{w} = \{w_1, \ldots, w_L\}$ in our training bitext[3] to establish initial correspondences between them:

$$P(\mathbf{c} \mid \mathbf{w}) = \sum_{\mathbf{a}} P(M) \prod_{j=1}^{M} P(a_j \mid L) P(c_j \mid a_j, w_{a_j}) \tag{3}$$

where $a_j \in \{0, \ldots, L\}, 1 \leq j \leq M$ are the alignment variables for the visual words c_j, and $a_j = 0$ is an assignment to the empty word. Once its parameters have been estimated, the model can be used to find the most likely translation:

$$\widetilde{\mathbf{w}} = \operatorname*{argmax}_{\mathbf{w}} P(\mathbf{w} \mid \mathbf{c}) = \operatorname*{argmax}_{\mathbf{w}} P(\mathbf{c} \mid \mathbf{w}) P(\mathbf{w}) \tag{4}$$

(Eq. 4) generates annotations $\widetilde{\mathbf{w}}$ when presented image features $\mathbf{c}$. The parameters of the translation model can be estimated from parallel data using an EM-style algorithm described by Brown et al. [4]. It simultaneously establishes explicit correspondences between instances of visual and caption words and estimates translation probabilities $t(c \mid w)$ for pairs of visual and textual words.

4 Compound Features

Despite good results for text translation tasks, the translation model described in Section 3 does not perform well when applied to boundary fragments and

[3] A body of two roughly aligned (e.g., on the level of sentences for bilingual texts) pieces of data.

caption words. There are several reasons for this problem: (i) individual boundary fragments are not discriminative enough to produce a reasonable low rate of false positive matches, and (ii) the translation model has difficulty in relating relatively many image features to relatively few caption words. As suggested by Wachsmuth et al. [11], Jamieson et al. [9] and others, this problem can be solved by grouping individual image features into larger compound objects. This approach has the additional benefit that such compounds can be augmented with descriptions of the spatial layouts of their component features in training images.

Definition 2. *A **compound-object** (or **compound** for short) is a pair* (L, R) *consisting of a set* $L = \{l_1, \ldots, l_n\}$ *of clusters and a set of sums of Gaussian densities* $R = \{d_{l_1 l_2} \mid l_1, l_2 \in L, l_1 \neq l_2\}$ *with* $d_{l_1 l_2} = \sum_{i=1}^{n_{l_1 l_2}} \mathcal{N}(\boldsymbol{\mu}_i, \sigma)$ *defining spatial relations between pairs of clusters.* $n_{l_1 l_2}$ *is determined by the procedure described at the end of this section. The vectors* $\boldsymbol{\mu}_i$ *determine the expected spatial offsets while the variance* σ *is fixed.*

The process of building compounds is driven by co-occurrences of boundary fragments in the training data. The method we use was suggested by Melamed [12] and is originally intended to identify Non-Compositional-Compounds (NCCs) like *hot dog* in bilingual texts.[4] Melamed's method is based around the observation that the mutual information

$$I(\mathcal{C}, \mathcal{W}) = \sum_{c \in \mathcal{C}} \sum_{w \in \mathcal{W}} P(c, w) \log \frac{P(c, w)}{P(c)P(w)} \tag{5}$$

of a translation model increases when the individual words that comprise NCCs are fused and treated as single tokens. The joint distribution $P(c, w)$ is determined by the alignments that the translation model produces for the visual and captions words of the training bitext. We use the following counting rule to obtain the joint distribution:

$$P(c, w) \propto \sum_{s=1}^{S} \underbrace{\min\{\#c^s, \#w^s\}}_{\text{number of links } c \to w} \overbrace{\frac{\min\{\#w^s, \#c^s_w\}}{\#c^s_w}}^{\text{weight}} \tag{6}$$

In (Eq. 6) the sum is over all pairs of the bitext, $\#c^s$, $\#w^s$ and $\#c^s_w$ refer to the number of instances of the respective token in the pair $(\mathbf{c}^s, \mathbf{w}^s)$, and c_w are visual words which translate to w. While in principle (Eq. 5), in conjunction with (Eq. 6), is sufficient to test any given NCC candidate using a trial translation model, the computational cost of this naive approach is not feasible for realistic numbers of NCC candidates. Therefore Melamed uses certain independence assumptions to derive a predictive value function V from (Eq. 5) that can be used to estimate the contributions of individual NCC candidates to the performance of the translation model without actually computing the mutual information:

[4] Melamed's approach was first applied to captionized datasets in [11]. However, results were only shown for synthetic symbol sets.

$$V(\mathcal{C},\mathcal{W}) = \sum_{c\in\mathcal{C}} V_{\mathcal{W}}(c) \qquad \text{where } V_{\mathcal{W}}(c) = P(c,w_c)\log\frac{P(c,w_c)}{P(c)P(w_c)} \tag{7}$$

Even though V allows us to test numerous NCC candidates c_1c_2 in parallel by independently computing their contributions $\Delta V_{c_1c_2}$, it is still too computationally expensive for practical applications. Fortunately further independence assumptions allow us to estimate the change of V caused by a NCC candidate without even introducing a trial translation model:

$$\begin{aligned}\widetilde{\Delta V}_{c_1c_2} = &\max\{V_{\mathcal{W}}(c_1 \mid s(c_1,c_2)), V_{\mathcal{W}}(c_2 \mid s(c_2,c_1)\} \\ &+ V_{\mathcal{W}}(c_1 \mid \neg s(c_1,c_2)) + V_{\mathcal{W}}(c_2 \mid \neg s(c_2,c_1)) - V_{\mathcal{W}}(c_1) - V_{\mathcal{W}}(c_2)\end{aligned}$$

The predicate s evaluates to true if an instance of the second argument is present in the image from which the instance of the first argument originates. The process of identifying compounds is very similar to Melamed's original method and consists of several iterations of the steps 2 – 8 of the following algorithm:

1. create empty lists of validated and rejected compounds
2. use bitext to train base translation model
3. produce candidate compounds for all pairs (c_1, c_2) that are not on list of rejected compounds and co-occur in at least two images of the training bitext and compute $\widetilde{\Delta V}_{c_1c_2}$ (in the first iteration, c_1 and c_2 are simply clusters; in later iterations, however, c_1 and c_2 can be compound objects themselves)
4. sort candidate list according to $\widetilde{\Delta V}_{c_1c_2}$ and discard candidates if (i) $\widetilde{\Delta V}_{c_1c_2} \leq 0$, or (ii) c_1 or c_2 is part of a candidate that is closer to the top of the list
5. train a test translation model in which components of candidate compounds are fused into single objects
6. discard and put on list of rejected compounds all candidates for which $\Delta V_{c_1c_2} \leq 0$
7. permanently fuse the remaining candidate compounds into single objects
8. goto step 2 until no more candidates are validated or the maximum number of iterations is reached

After compounds have been identified, information about typical spatial relations of the components of the compounds are added. This is done by finding images in which boundary fragments of different clusters of a compound are present. Then relative positions of the boundary fragments give rise to spatial relations. Relative positions of boundary fragments in training images are shown in Fig. 2. Finally, optimal values for the individual detection thresholds of all clusters are computed using the translation model and the training images. As each compound possesses a word as its most likely translation, it is known in which training images a given compound should be detected. Therefore the detection threshold h can be adjusted to ensure that clusters are detected in positive but not in negative training images. In order to avoid overfitting, distorted versions of the training images are used in this process.

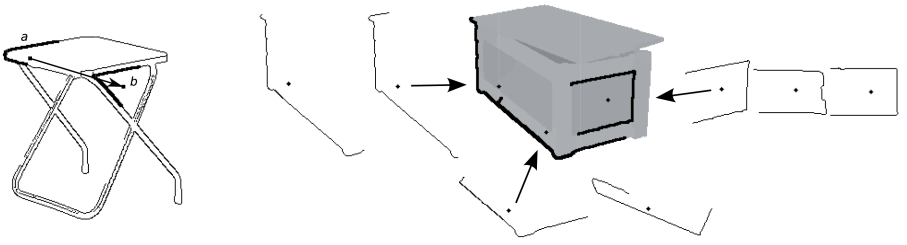


Fig. 2. Relative positions of boundary fragments. The left image shows the extraction of a spatial relation from a training image, the right image visualizes a compound generated from the furniture dataset. The different fragments of each component cluster are shown.

5 Automatic Image Annotation

The most obvious application[5] of the correspondence data created by the presented process is automatic image annotation, i.e., assigning suitable caption words to previously unseen and uncaptioned images. In our approach, automatic annotation is done by detecting compounds in the presented images and assigning the most likely translation words of the detected compounds as caption words. However, since detecting boundary fragments always produces localized results, it is also possible to relate image regions to caption words, therefore allowing region naming as an extension of automatic annotation.

When presented an unknown image, in principle, all boundary fragments could be matched against the image using chamfer matching, which would lead to zero or more matching locations for each boundary fragment. However, this approach is not feasible because of the large number of chamfer matching operations necessary. Instead we only match boundary fragments previously chosen to represent clusters in order to quickly rule out clusters of boundary fragments that probably do not match the image. The whole annotation procedure works as follows:

1. obtain edge image from input image
2. find matching boundary fragments using the following two-step approach
   (a) for each cluster (F, r, h): match the representative boundary fragment r against the image; discard cluster when edge distance is above a global coarse threshold H
   (b) match all boundary fragments F of the remaining clusters against the image; store locations of all matches for which the edge distance is below the cluster-specific tight threshold h; discard clusters that do not produce at least one match
3. for all compounds for which all component clusters have been detected in the image in at least one position: check spatial relations for all possible

[5] For a more extended list of applications of multimodal correspondence models, see Duygulu et al. [2].

combinations of detected cluster instances (each of which could comprise a compound instance); store combination that fulfills spatial relations best; discard compound if such a combination does not exist
4. output most likely translation words of remaining compounds as caption

6 Results

In order to evaluate the annotation performance, we used our method to automatically annotate training images and held-out test images from two datasets. As both datasets contained full annotations, it was possible to compare original and generated caption words. As in the evaluations of comparable methods, we use precision and recall to quantify annotation performance.

The first dataset we evaluated was obtained from the product catalogue of a large European supplier of furniture by extracting product images and head nouns of the related product descriptions. The resulting dataset consists of images of single pieces of furniture with simple backgrounds along with sets of several caption words that mostly refer to the depicted objects. The 525 image-caption pairs of the dataset were divided into 300 training pairs and 225 test pairs. The dataset features some challenging aspects, namely large within-category shape variations, different words for similarly shaped pieces of furniture, and a number of categories that occur very rarely. Fig. 3 shows three examples of automatic annotations. The precision and recall values for automatic annotations of the training and test subsets of the furniture dataset are shown in Fig. 4. The low precision values for most object categories were caused by clusters of shapes (like legs, see Fig. 1) that occur in a wide variety of object categories.

The second dataset used in the evaluation was synthetically generated to allow fine grained control over shape variation and image clutter. As the first dataset turned out to be quite challenging, the generated dataset was tuned to be simpler in some regards. A probabilistic grammar was developed to construct simple hierarchically structured objects while allowing variations in structure and shape. We distinguished five different object categories, some of which shared common elementary shapes. For the training, suitable caption words were emitted along with an adjustable number of random words. All words were chosen from a pool of 27 words. The generated pairs were divided into 300 training and 200 test pairs. For test images, which were used in the annotation task, random

wall, uplighter, **lamp**, **lamp**, bookcase, chair, floor

table, beech, **clock**, **clock**

bookcase, wall, lamp, holder, uplighter, bookcase, **chair**, beech, **chair**, ...

Fig. 3. Results of automatic annotation for furniture dataset. Generated annotation words are listed on the right side of the respective test image, correct annotation words appear in bold print.

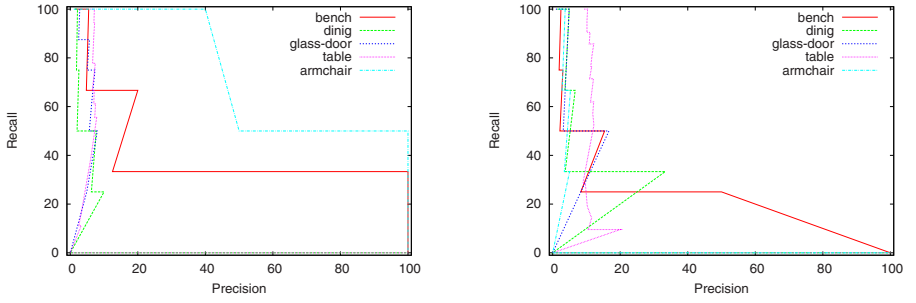
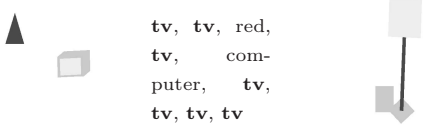


Fig. 4. Precision vs. recall for furniture dataset. Each curve shows precision and recall for a single concept. Left: training subset. Right: test subset.



tv, **tv**, red, **tv**, computer, **tv**, **tv**, **tv**, **tv**

combination, **lamp**, **lamp**, **lamp**, kitchen, **lamp**, white, **lamp**, **lamp**, **lamp**, **lamp**, wall-mounted, kitchen, small, conference

Fig. 5. Results of automatic annotation for synthetic dataset. Generated annotation words are listed on the right side of the respective test image, correct annotation words appear in bold print.

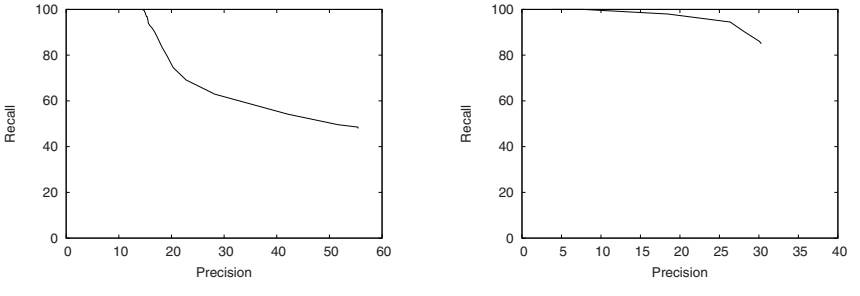


Fig. 6. Precision vs. recall for synthetic dataset. Left: training subset. Right: test subset.

shapes were added to emulate background clutter. As the synthetically generated objects were treated as regular images, the algorithm also needed to deal with occlusion as well as over- and under-segmentation issues. As results show (see Fig. 6), our method produced acceptable precision and recall when applied to the training portion of the synthetic dataset. For the test portion however, the achievable precision was considerably lower. Fig. 5 shows two examples of the generated annotations.

7 Conclusion

In this paper, we presented on-going work on learning visual compounds from image-caption pairs. We focused on shape-based feature descriptions that provide a suitable characterization for many categories of man-made objects. The approach is able to deal with slight occlusion as well as moderate over- and under-segmentation by using boundary fragments for building a basic visual vocabulary. The grouping of fragments into compounds is driven by caption words using a translation model and a mutual information measure. We have shown that the approach works on a set of synthetically generated images with captions. First tests on a realistic dataset from a furniture catalogue showed the generation of some promising compounds but also indicated several problems regarding the discriminability of boundary fragments and the sparseness of this kind of dataset.

References

1. Barnard, K., Duygulu, P., Forsyth, D., de Freitas, N., Blei, D., Jordan, M.: Matching Words and Pictures. Journal of Machine Learning Research 3, 1107–1135 (2003)
2. Duygulu, P., Barnard, K., de Freitas, J.F.G., Forsyth, D.A.: Object Recognition as Machine Translation: Learning a Lexicon for a Fixed Image Vocabulary. In: Proc. of 7th Europ. Conf. on Computer Vision, Copenhagen, vol. 4, pp. 97–112 (2002)
3. Hofmann, T.: Learning and representing topic. A hierarchical mixture model for word occurrence in document databases. In: Proc. Workshop on learning from text and the web, CMU (1998)
4. Brown, P.F., Pietra, S.A.D., Mercer, R.L., Pietra, V.J.D.: The Mathematics of Statistical Machine Translation: Parameter Estimation. Computational Linguistics 19(2), 263–311 (1993)
5. Berg, T., Berg, A., Edwards, J., Maire, M., White, R., Teh, Y.W., Learned-Miller, E., Forsyth, D.: Names and faces in the news. In: Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR) (2004)
6. Carneiro, G., Chan, A.B., Moreno, P.J., Vasconcelos, N.: Supervised Learning of Semantic Classes for Image Annotation and Retrieval. IEEE Transactions on Pattern Analysis and Machine Intelligence 29(3), 394–410 (2007)
7. Crandall, D., Huttenlocher, D.: Weakly supervised learning of part-based spatial models for visual object recognition. In: Leonardis, A., Bischof, H., Pinz, A. (eds.) ECCV 2006. LNCS, vol. 3954. Springer, Heidelberg (2006)
8. Opelt, A., Pinz, A., Zisserman, A.: A Boundary-Fragment-Model for Object Detection. In: Leonardis, A., Bischof, H., Pinz, A. (eds.) ECCV 2006. LNCS, vol. 3954. Springer, Heidelberg (2006)
9. Jamieson, M., Dickinson, S., Stevenson, S., Wachsmuth, S.: Using Language to Drive the Perceptual Grouping of Local Image Features. In: Proceedings of the 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2006), New York, vol. 2, pp. 2102–2109 (2006)
10. Jamieson, M., Fazly, A., Dickinson, S., Stevenson, S., Wachsmuth, S.: Learning Structured Appearance Models from Captioned Images of Cluttered Scenes. In: Proc. of the Int. Conf. on Computer Vision (ICCV), Rio de Janeiro (October 2007)

11. Wachsmuth, S., Stevenson, S., Dickinson, S.: Towards a Framework for Learning Structured Shape Models from Text-Annotated Images. In: Proceedings of the HLT-NAACL 2003 workshop on Learning word meaning from non-linguistic data, Edmonton, vol. 6, pp. 22–29 (2003)
12. Melamed, I.D.: Automatic Discovery of Non-Compositional Compounds in Parallel Data. In: Proceedings of the 2nd Conference on Empirical Methods in Natural Language Processing, Providence, pp. 97–108 (1997)
13. Christoudias, C.M., Georgescu, B., Meer, P.: Synergism in Low Level Vision. In: 16th Int. Conf. on Pattern Recognition, Quebec City, vol. 4, pp. 150–155 (2002)
14. Borgefors, G.: Hierarchical chamfer matching: A parametric edge matching algorithm. Trans. on Pattern Analysis and Machine Intelligence 10, 849–865 (1988)