# Hierarchical Learning of Dominant Constellations for Object Class Recognition

Nathan Mekuz and John K. Tsotsos

Center for Vision Research (CVR) and
Department of Computer Science and Engineering,
York University, Toronto, Canada M3J 1P3
{mekuz,tsotsos}@cse.yorku.ca

**Abstract.** The importance of spatial configuration information for object class recognition is widely recognized. Single isolated local appearance codes are often ambiguous. On the other hand, object classes are often characterized by groups of local features appearing in a specific spatial structure. Learning these structures can provide additional discriminant cues and boost recognition performance. However, the problem of learning such features automatically from raw images remains largely uninvestigated. In contrast to previous approaches which require accurate localization and segmentation of objects to learn spatial information, we propose learning by hierarchical voting to identify frequently occurring spatial relationships among local features directly from raw images. The method is resistant to common geometric perturbations in both the training and test data. We describe a novel representation developed to this end and present experimental results that validate its efficacy by demonstrating the improvement in class recognition results realized by including the additional learned information.

## 1 Introduction

Humans are highly adept at classifying and recognizing objects with significant variations in shape and pose. However, the complexity and degree of variance involved make this task extremely challenging for machines. Current leading edge methods use a variety of tools including local features [1,2,3,4], global [5] and region [6] histograms, dominant colors [7], textons [8] and others, collecting features sparsely at detected key points, at random locations, or densely on a grid and at single or multiple scales. In practice, different types of features are often complementary and work well in different scenarios, and good results are often achieved by combining different classifiers. Of the above approaches, much focus has recently been dedicated to learning with local appearance descriptors, which have been shown to be extremely effective thanks to their discriminant qualities and high degree of resistance to geometric and photometric variations as well as partial occlusions.

A very effective and widely-used technique that enables the use of efficient search methods borrowed from the text retrieval field is vector quantization,

Y. Yagi et al. (Eds.): ACCV 2007, Part I, LNCS 4843, pp. 492–501, 2007.

whereby each patch is associated with a label (visual word) from a vocabulary. The vocabulary is usually constructed offline by means of some clustering algorithm. To avoid aliasing effects arising from boundary conditions, soft voting is employed, whereby each vote is distributed into several nearby words using some kernel function. Finally, images are coded as histograms of their constituent visual words.

While the importance of local features' spatial configuration information for object class recognition is widely recognized, the basic scheme described above is typically employed on sets of isolated local appearance descriptors. However, for the most part, local appearance descriptors were designed to recognize local patches. When used for recognizing objects, the spatial layout that they appear in is of paramount importance. The SIFT algorithm [9], for example, represents local features in a way that is invariant to geometric perturbations. However, it also stores the parameters of the local geometry, and subsequently applies a Hough transform to select from potential hypotheses a model pose that conforms to the geometry associated with a large number of identified keys.

Current systems that capture spatial information do so by learning and enforcing local relationships [10,11], global relationships [12,13], using dense sampling [2,1], or at multiple levels [14,15]. In [14], the system learns groups of local features that appear frequently in the training data, followed by global features composed of local groupings. In [11], spatial consistency is enforced by requiring a minimum number of features to co-occur in a feature neighborhood of fixed size. The authors of [16,12] demonstrate the benefit of learning the spatial relationships between various components in an image from a vocabulary of relative relationships. In [2], appearance models are built where clusters are learned around object centers and the object representation encodes the position and scale of local parts within each cluster. Significant performance gains are reported resulting from the inclusion of location distribution information. Fergus et al. [1] learn a scale-normalized shape descriptor for localized objects. However, the shapes are not normalized with respect to any anchor point. Consequently, some preprocessing of the input images is required. A boosting algorithm that combines local feature descriptors and global shape descriptors is presented in [13], however extracting global shape is extremely difficult under occlusion or cluttered background.

We take a different approach and seek to learn object class-specific hierarchies of constellations, based on the following principles:

**Unsupervised learning.** A clear tradeoff exists between the amount of training data required for effective learning, and the quality of its labeling. Given the high cost of manual annotation and segmentation, and the increased availability (e.g. on the internet) of images that are only globally annotated with a binary class label, a logical goal is the automatic learning of constellation information from images with minimal human intervention. Specifically, this precludes manual segmentation and localization of objects in the scene.

**Invariance to shift, scale and rotation.** In order to be able to train with and recognize objects in various poses, we require a representation that

captures spatial information, yet is resistant to common geometric perturbations.

**Robustness.** In order to successfully learn in an unsupervised fashion, the algorithm must be robust to feature distortions and partial occlusion. A common approach for achieving robustness is voting.

**Learn with no spatial restrictions** We would like to learn spatial relationships over the entire image, without restrictions of region or prior (e.g. Gestalt principles). This allows grouping discontinuous features, e.g. such that lie on the outline of an object with a variable texture interior.

The main contribution of this paper is a novel representation that captures spatial relationship information in a scale and rotation invariant manner. The constellation descriptors are made invariant by anchoring with respect to one local feature descriptor, similar to the way the SIFT local descriptor anchors with respect to the dominant orientation. We present a framework for learning spatial configuration information by collecting inter-patch statistics hierarchically in an unsupervised manner. To tackle the combinatorial complexity problem, higher level histograms are constructed by successive pruning. The most frequently occurring constellations are learned and added to the vocabulary as new visual words. We also describe an efficient representation for matching learned constellations in novel images for the purpose of object class recognition or computing similarity.

The remainder of this paper is organized as follows: in Section 2 we describe our proposed constellation representation. This is followed by implementation details of the voting scheme in Section 3, and the matching algorithm in Section 4. The results obtained on images of various categories are presented in Section 5, and finally, Section 6 concludes with a discussion.

## 2 Invariant Constellation Representation

The constellation representation captures the types and relative positions, orientations and scales of the constituent parts. An effective representation must

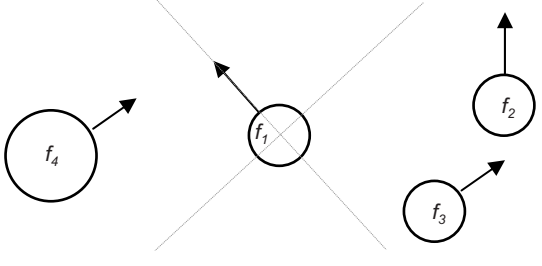


**Fig. 1.** An illustration of the constellation representation. Local features are represented with circles, with arrows emanating out of them to indicate dominant orientation. Feature $f_1$ is selected as anchor, and the positions, scales and orientations of the remaining features are expressed relative to it.

be resistant to minor distortions arising from changes in pose or artifacts of the local feature extraction process. Pose changes can have a significant effect on the coordinates of local features.

Another key requirement is a consistent frame of reference. In the absence of models of localized objects, a frame of reference can be constructed as a function of the constituent features. One option is to use the average attributes of local features [17]. However, since our method uses local features that are quantized into discrete visual words, we opt for the simpler alternative of pivoting at the feature with the lowest vocabulary index. This results in a more compact representation (and in turn, computational complexity savings) by eliminating the need to store spatial information for the anchor feature. On the downside, the anchor feature may not lie close to the geometric center of the constellation, reducing the granularity of position information for the other features. Whatever method is used for selecting the pivot, detection of the constellation depends on the reliable recovery of the pivot feature. However, even if the pivot feature cannot be recovered (e.g. under occlusion), subsets of the constellation may still be detected.

Our representation is illustrated graphically in Figure 1, with the local appearance features represented as circles, and their dominant orientation as arrows. Feature $f_1$ is selected as anchor and the coordinate system representing the remaining constellation features is centered about it and rotated to align with its dominant orientation. More formally, given a set of local features $\mathcal{F}_i = \langle \Gamma_i, t_i, x_i, y_i, \alpha_i \rangle$ where $\Gamma_i$ is the index of the visual word corresponding to feature $\mathcal{F}_i$, $t_i$ is its scale, $x_i$ and $y_i$ its position in the image and $\alpha_i$ its orientation, relative to the global image coordinate system, we select the anchor $\mathcal{F}_*$ as $\mathcal{F}_* = \arg\min \Gamma_i$, and construct the constellation descriptor encoding the anchor feature's type $\Gamma_*$, as well as the following attributes for each **remaining** feature $\mathcal{F}_j$:

**Type:** $\Gamma_j$

**Scale ratio:** $t_j/t_*$

**Relative orientation:** $\alpha_j - \alpha_*$

**Relative position:** $\text{atan2}(y_j - y_*, x_j - x_*) - \alpha_*$ where atan2 is the quadrant-sensitive arctangent function. This attribute ignores distances, and merely provides a measure of $\mathcal{F}_j$'s polar angle relative to $\mathcal{F}_*$, using $\mathcal{F}_*$'s coordinate system.

As an example, using this representation, a pair of local features $\{\mathcal{F}_1, \mathcal{F}_2\}$ with $\Gamma_1 < \Gamma_2$ is represented as $\langle \Gamma_1, \Gamma_2, t_2/t_1, \alpha_2 - \alpha_1, \text{atan2}(y_2 - y_1, x_2 - x_1) - \alpha_1 \rangle$. A constellation of $m$ local features is represented as an $n$-tuple with $n = 4m - 3$ elements. To maintain consistent representation, the descriptor orders the local features by their vocabulary index. If the lowest vocabulary index is not unique to one local feature, we build multiple descriptors, just as the SIFT algorithm creates multiple descriptors at each keypoint where multiple dominant orientations exist.

## 3 Voting by Successive Pruning

The learning phase performs histogram voting in order to identify the most frequently occurring constellations in each category, using the representation described above. Since the descriptor orders the local features by their type attribute, each resulting histogram takes the shape of a triangular hyper-prism, with the $\Gamma$ (type) axes along the hyper-triangular bases and the other attributes forming the rectangular component of the prism. Spatial information is encoded in $8 \times 8 \times 8$ bins. The relative orientation and relative position attributes encode the angle into one of eight bins, similar to way this is done in SIFT. Scale ratios are also placed into a bin according to $\log_2(t_j/t_*) + 3$. Co-occurrences with scale ratios outside the range $[1/16, \ldots, 32]$ are discarded. As in SIFT, in order to avoid aliasing effects due to boundary conditions, we use soft voting whereby for each attribute, each vote is hashed proportionally into two neighboring bins. For the type attribute, the vote is distributed into several nearby visual words using a kernel. We use a Gaussian kernel with $\sigma$ set to the average cluster radius, although other weighting formulas are certainly possible. It is also possible to threshold by distance rather than fixing the number of neighbors.

The exponential complexity of the problem, and in particular, the size of the voting space, call for an approximate solution. A simple method that has often been used successfully is successive pruning. In computer vision, good results have been reported in [18] although some human supervision was necessary at the highest levels of the hierarchy. In some sense, the successive pruning strategy can be viewed as a coarse-to-fine refinement process. Starting with coarse bins, the algorithm identifies areas of the search space with a high number of counts. It then iteratively discards bins with a low number of counts, re-divides the rest of the voting space into finer bins, and repeats the voting process. In the case of multi-dimensional histograms, coarse bins can also be created by collapsing dimensions. This latter approach is more convenient in our case since it fits

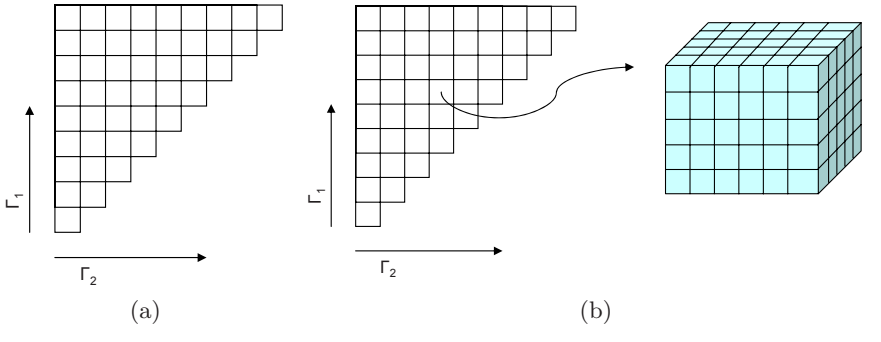


**Fig. 2.** (a) A depiction of a triangular histogram used for voting for the most frequently co-occurring pairs. (b) Histogram pruning: the bin with a low number of counts are discarded. Finer bins are allocated for each bin with a number of counts above a threshold.

naturally with the notion of hierarchical learning, creating larger constellations from smaller ones. Also, collapsing the dimensions associated with spatial information offers computational advantages by allowing early termination of the voting in the discarded bins, since visual word indices for local features are available immediately.

Figures 2(a) and 2(b) illustrate the structures used in the two-phase voting process to identify pairs of local features that appear frequently in a particular spatial configuration. In the first phase, local feature descriptors are extracted and cached from all images belonging to an object class, and a triangular histogram (Fig. 2(a)) is collected to count the number of times each pairs of local features co-occurs, regardless of geometry. In the second phase (Fig. 2(b)), sub-histograms are allocated for the bins with a high number of counts, and the remaining bins are discarded. Each sub-histogram consists of $8 \times 8 \times 8$ bins and captures spatial information for its associated bin in the triangular histogram. Finally, the vocabulary is augmented with new visual words corresponding to the most frequent constellations in each class identified in phase 2.

## 4 Indexing Constellation Descriptors for Efficient Matching

Given novel input images, the system compares constellations extracted from these images against the learned constellation stored in its vocabulary. In order to achieve this, an exhaustive search of all local feature combinations in the input images is not necessary. A more efficient search is possible by indexing the learned constellation information offline, as depicted in Figure 3. At the first level, the structure consists of a single array indexed by local feature type index. Given a moderate number of learned constellations, the resulting first level array is sparse. Local feature types for which learned constellations exist, have their array entries point to arrays of stored constellation descriptors, sorted lexicographically by the other type indices.

The matching algorithm works by constructing an inverted file [19] of the local features in the image. A sparse inverted file containing only links to vocabulary

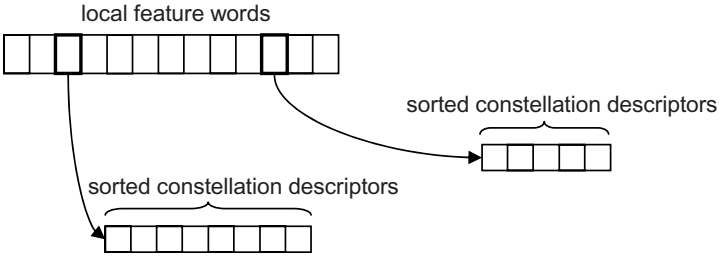


**Fig. 3.** A depiction of an efficient indexing structure for fast lookup of constellation features

entries that are in use suffices thanks to the sorted second-level arrays: a match is sought simply by traversing both lists simultaneously. Spatial relationship information is compared only when all type attributes in a constellation descriptor are matched in the inverted file.

## 5 Evaluation

We tested our technique by examining the effect of using spatial relationship information captured using our descriptors on object class recognition performance. In order to isolate the effect of our constellation learning algorithm on class detection, we limited our evaluation to still greyscale images. We used the SIFT detector and descriptor to extract and represent local appearance features. We constructed a vocabulary of 13,000 visual words by extracting features from the first 800 hits returned by Google Images for the keyword 'the' and clustering with k-means. The result is a *neutral* vocabulary, that is not tuned specifically for any object category. For training and test data, we used 600 images of faces, airplanes, watches, bonsai trees and motorbikes from the PASCAL 2006 data set [20], divided equally into training and test images. All images were converted to greyscale, but no other processing was performed.

In the training phase, image descriptors were collected for each of the training images encoding histograms of their constituent visual words. We used our neutral vocabulary constructed as described above, and quantized each feature descriptor to its 15 nearest neighbors using a Gaussian kernel with $\sigma$ set to the average radius covered by each vocabulary entry. In the testing phase, each image was matched and classified using a simple unweighted nearest neighbor classifier against the trained image descriptors. As is standard practice, we used

| | Faces | Airplanes | Watches | Bonsai trees | Motorcycles |
|---|---|---|---|---|---|
| Faces | **88.3** | 3.3 | 5.0 | 1.7 | 1.7 |
| Airplanes | 53.3 | **36.7** | 0.0 | 8.3 | 1.7 |
| Watches | 45.0 | 11.7 | **26.7** | 16.7 | 0.0 |
| Bonsai trees | 21.7 | 1.7 | 1.7 | **71.7** | 3.3 |
| Motorcycles | 70.0 | 5.0 | 1.7 | 8.3 | **15.0** |

(a)

| | Faces | Airplanes | Watches | Bonsai trees | Motorcycles |
|---|---|---|---|---|---|
| Faces | **91.7** | 3.3 | 5.0 | 0.0 | 0.0 |
| Airplanes | 50.0 | **43.3** | 0.0 | 5.0 | 1.7 |
| Watches | 40.0 | 10.0 | **35.0** | 13.3 | 1.7 |
| Bonsai trees | 16.7 | 3.3 | 1.7 | **76.7** | 1.7 |
| Motorcycles | 66.7 | 5.0 | 0.0 | 6.7 | **21.7** |

(b)

**Fig. 4.** Confusion matrices (a) using a vocabulary of only local appearance features. (b) using an augmented vocabulary with an additional 50 constellation words per category.

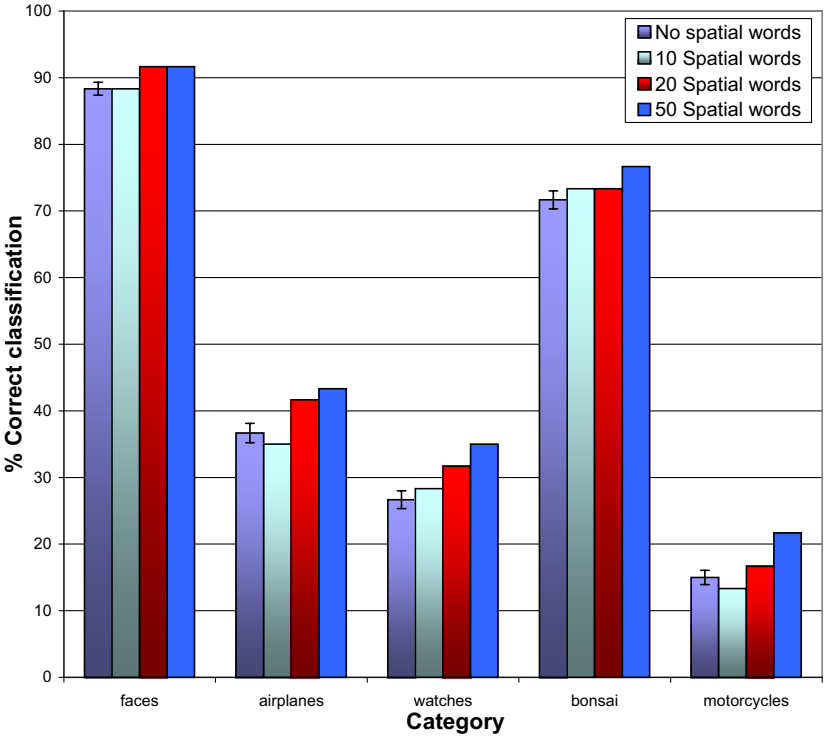


**Fig. 5.** Object class categorization performance as a function of the number of constellation visual words used. The error bars represent a margin of 3 standard errors.

a stop list to discard the 2% most frequently occurring visual words. Although more elaborate classifiers (e.g. SVM) and weighting schemes (e.g. tf-idf [21]) are possible, we opted for the simple scheme described here in order to focus on the effect of the additional spatial information. We expect the tf-idf scheme to place increased weights on the constellation features, since they carry more class-specific discriminant information.

We tested the effect of augmenting the vocabulary with 10, 20 and 50 constellation words per object class on class recognition performance. It is worth noting that the vocabulary used for constructing these additional constellation words was again our generic neutral vocabulary: the only class-specific information captured in the training phase was the most frequently co-occurring pairs and their spatial relationships in each class. A 2% stop list was again used on the visual words associated with the local features but not on the pairs.

Figure 4 presents confusion matrices for the categorization tests (a) using no spatial information, and (b) using 50 additional constellation visual words. Perhaps surprisingly, poor performance is realized in the motorcycles category, where local feature-based methods typically excel. A likely explanation is that normally the vocabulary is constructed using images of the modeled class, and

captures features such as wheels in the case of the motorcycle class, whereas in our experiments we used a neutral vocabulary that was not trained specifically for any class. More importantly, however, we note that the addition of a few visual words corresponding to learned spatial features clearly boosted recognition performance in all classes, with average gains of about 5% using 50 constellation words. Figure 5 shows correct recognition results (corresponding to the diagonal of the confusion matrices) with different numbers of constellation words. The general trend shows recognition performance improving as more constellation features are used. The error bars represent an interval of 3 standard errors.

## 6 Discussion

This paper has presented a novel approach for representing constellation information that is learned directly from raw image data in a hierarchical fashion. The method is capable of learning spatial configuration information from possibly cluttered images where objects appear in various poses and possibly partly occluded. Novel images are tested for the presence of learned configurations in a way that is robust to common geometric perturbations. Additionally, the paper presents implementation details for an efficient voting algorithm that allows collecting robust co-occurrence statistics in a computationally highly complex voting space, and efficient indexing structures that allow fast lookup in the matching phase. Our experimental results confirm the importance of spatial structure to the class recognition problem, and show that the proposed representation can provide significant benefit with constellations consisting of pairs. We are currently exploring richer constellation structures corresponding to higher levels of the hierarchy and looking at ways for visualizing the learned constellations.

## Acknowledgments

The authors are grateful to Erich Leung and Kosta Derpanis for many helpful discussions. This work was supported by OGSST and Precarn incorporated.

## References

1. Fergus, R., Perona, P., Zisserman, A.: Object class recognition by unsupervised scale-invariant learning. In: IEEE Conference on Computer Vision and Pattern Recognition, vol. 2, pp. 264–271 (2003)
2. Leibe, B., Mikolajczyk, K., Schiele, B.: Efficient clustering and matching for object class recognition. In: British Machine Vision Conference, Edinburgh, England (2006)
3. Berg, A.C., Berg, T.L., Malik, J.: Shape matching and object recognition using low distortion correspondences. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 26–33. IEEE Computer Society Press, Los Alamitos (2005)
4. Dorko, G., Schmid, C.: Object class recognition using discriminative local features (2005)

5. Ortega, M., Rui, Y., Chakrabarti, K., Mehrotra, S., Huang, T.S.: Supporting similarity queries in mars. In: ACM International Conference on Multimedia, pp. 403–413. ACM Press, New York (1997)
6. Carson, C., Thomas, M., Belongie, S., Hellerstein, J., Malik, J.: Blobworld: a system for region-based image indexing and retrieval. Technical report, Berkeley, CA, USA (1999)
7. Mukherjea, S., Hirata, K., Hara, Y.: Amore: a world-wide web image retrieval engine. In: CHI 1999. Extended abstracts on human factors in computing systems, pp. 17–18. ACM Press, New York (1999)
8. Malik, J., Belongie, S., Shi, J., Leung, T.K.: Textons, contours and regions: Cue integration in image segmentation. In: IEEE International Conference on Computer Vision, pp. 918–925. IEEE Computer Society Press, Los Alamitos (1999)
9. Lowe, D.G.: Object recognition from local scale-invariant features. In: IEEE International Conference on Computer Vision, vol. 1150, IEEE Computer Society Press, Los Alamitos (1999)
10. Lazebnik, S., Schmid, C., Ponce, J.: Affine-invariant local descriptors and neighborhood statistics for texture recognition. In: IEEE International Conference on Computer Vision, vol. 649, IEEE Computer Society, Los Alamitos (2003)
11. Sivic, J., Zisserman, A.: Video Google: A text retrieval approach to object matching in videos. In: IEEE International Conference on Computer Vision, vol. 2, pp. 1470–1477 (2003)
12. Lipson, P., Grimson, E., Sinha, P.: Configuration based scene classification and image indexing. In: IEEE Conference on Computer Vision and Pattern Recognition, vol. 1007, IEEE Computer Society, Los Alamitos (1997)
13. Zhang, W., Yu, B., Zelinsky, G.J., Samaras, D.: Object class recognition using multiple layer boosting with heterogeneous features. In: IEEE Conference on Computer Vision and Pattern Recognition
14. Amit, Y., Geman, D.: A computational model for visual selection. Neural Comput. 11, 1691–1715 (1999)
15. Agarwal, A., Triggs, W.: Hyperfeatures - multilevel local coding for visual recognition. In: Leonardis, A., Bischof, H., Pinz, A. (eds.) ECCV 2006. LNCS, Springer, Heidelberg (2006)
16. Sinha, P.: Image invariants for object recognition. Invest. Opth. & Vis. Sci. 34(6) (1994)
17. Shokoufandeh, A., Dickinson, S.J., Jönsson, C., Bretzner, L., Lindeberg, T.: On the representation and matching of qualitative shape at multiple scales. In: European Conference on Computer Vision, pp. 759–775. Springer, Heidelberg (2002)
18. Fidler, S., Berginc, G., Leonardis, A.: Hierarchical statistical learning of generic parts of object structure. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 182–189. IEEE Computer Society Press, Los Alamitos (2006)
19. Witten, I.H., Moffat, A., Bell, T.C.: Managing gigabytes: compressing and indexing documents and images, 2nd edn. Morgan Kaufmann Publishers Inc, San Francisco (1999)
20. Everingham, M., Zisserman, A., Williams, C.K.I., Van Gool, L.: The PASCAL Visual Object Classes Challenge. In: VOC2006 (2006)
21. Baeza-Yates, R., Ribeiro-Neto, B.: Modern Information Retrieval. Addison-Wesley, Reading (1999)