Binary Co-occurrences of Weak Descriptors

Martin Winter and Horst Bischof
Graz University of Technology, Austria [*]

Abstract

This paper demonstrates that a reliable and efficient object recognition system based only on binary joint occurrences of quantized descriptors can be built. Specifically, we show that a high recognition performance can be obtained even with very weak (non discriminative) descriptors. The binary joint occurrence representation despite being high dimensional is very sparse and therefore efficient. In order to obtain reliable joint occurrences we present a fast hierarchical quantization algorithm. We illustrate our results using different descriptors (PCA-SIFT, Spin images, SIFT) on a challenging, specific object recognition task and consider the scaling behavior of the method.

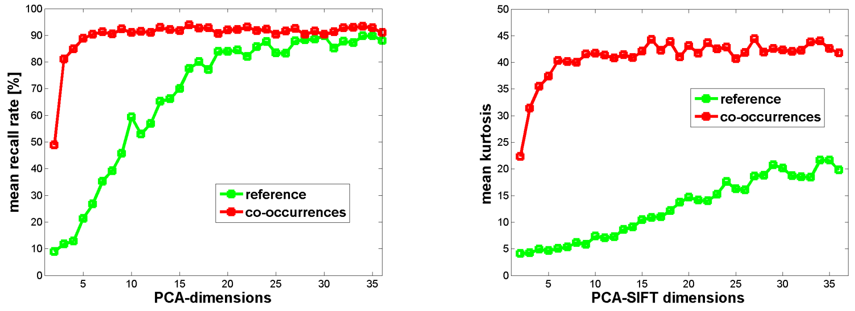
1 Introduction

In the last decade, object recognition tasks based on local features gained more and more interest by the computer vision community. A lot of different approaches have been proposed and recent evaluations have shown satisfactory performance on specific as well as generic object recognition challenges (*e.g.* [2, 5, 8, 16]). To increase the discriminative power of the approaches, the algorithms tend to be more and more complex thus requiring increasing computational power, runtime and necessary amount of memory. For example, key-point detectors are made robust against image distortions, viewpoint- and illumination changes [14, 15], descriptors are driven toward enhanced discriminative power, distinctiveness and invariance [12, 14] and complex selection and decision algorithms such as support vector machines or boosting algorithms [6, 23] have been applied to several recognition systems. It is evident that approaches considering single key-points and descriptors (bag of key-point approaches such as *e.g.* [5]) have severe limitations by not taking spatial neighborhood relations into account. Therefore, recent work has investigated spatial relations among key-points. Generative models use the spatial relations directly in modeling the assembly of object parts. One prominent example is the 'constellation model' of Fergus *et al.* [9], a fully connected probabilistic model of a few object parts trained to identify object classes on unsegmented, cluttered scenes. Simpler 'relational models' have been proposed by *e.g.* Crandall and Huttenlocher (k-fan model) [4] or Fergus *et al.* (star-shape model) [10]. Carneiro and Lowe [3] proposed an approach which does not rely on strict spatial models, but allows to adapt the spatial relations specifically to the underlying object properties. Another important approach, which primarily tries

[*] This work has been funded by the Austrian Joint Research Project Cognitive Vision under sub-projects S9103-N04 and S9104-N04 and it has been partially supported by the European Union Network of Excellence, MUSCLE, under contract FP6-507752.

BMVC 2007 doi:10.5244/C.21.13

to overcome the limitations on the number of distinct features in the model, is the one from Bouchard and Triggs [2], where the authors proposed a probabilistic, hierarchical approach on features and spatial relations of object parts. Another class of approaches uses spatial relations as additional information. Those spatial relations are either used to improve the discriminative power of features directly (*e.g.* ‘Hyperfeatures’ [1] by Agarwal and Triggs) or they are used for verification of tentative matches (*e.g.* ‘semi-local constraints’ by Schmid and Mohr [21]). Another example for spatial relations in a second processing step is the texture recognition system proposed by Lazebnik *et al.* [13] where spatial neighborhood statistic is used for final texture classification. There are also other interesting approaches that use spatial relations such as the one from Sivic *et al.* [7] (classification with probabilistic Latent Semantic Analysis (pLSA) in a ‘bag of words’ model) or the joint spatial relation of boundary fragments from Opelt *et al.* [19] . In summary all these papers (and many others) demonstrate, that spatial relations can significantly improve recognition results.



(a) (b)

Figure 1: Co-occurrence and reference algorithm mean recall rates (a). Mean kurtosis values of our object recognition task for decreasing PCA-SIFT descriptor size (b).

This evidence triggered the work in this paper where we take an extreme case. The only information we consider is the joint occurrence or co-occurrence of quantized descriptors. Moreover, we only represent the presence or absence of these co-occurrences which leads to a binary representation. The intuition behind is, that the co-occurrence of descriptors is a very discriminative feature because it is very unlikely, that two descriptors co-occur just by chance. In particular, we show that due to the distinctiveness of co-occurrences we can work even with very weak descriptors which are not discriminative by themselves. The following experiment demonstrates this fact (full details are explained in the experimental section 3). We use a straight forward, state of the art recognition setup: key-point extraction, descriptor calculation, quantization and matching. As descriptor we selected PCA-SIFT [12] proposed by Ke and Sukthankar, because we can control the weakness of that descriptor by consecutively reducing the descriptor dimension. The main result of the experiment is depicted in Figure 1a and shows the mean recall rate as function of the diminishing information (PCA-dimension). First one sees as expected, that using a single descriptor decreasing the PCA dimension decreases the recognition rate. More important, using co-occurrences of descriptors the recognition rate remains high despite the increasing weakness of the descriptor. This indicates, that the co-occurrence approach is not only much more discriminative, but can deal with ‘weaker’ descriptors still preserving a high recognition rate. In the following section we present the details of the binary co-occurrence approach and the proposed recognition system. In the experimental section 3 we describe a set of experiments which foster our assumptions, show that co-occurrences also increase the reliability and significance of an object recognition task and consider the scaling behavior.

2 Binary Co-occurrence overview

Our system has four main components, namely: 1. identification of key-points and calculation of a descriptor, 2. clustering the descriptors of the training database features to obtain a visual vocabulary, 3. nearest neighbor assignment of features and building the binary co-occurrence matrix representation and 4. matching the co-occurrence matrices against the database representation. An overview of the components and their interaction is depicted in Figure 2.

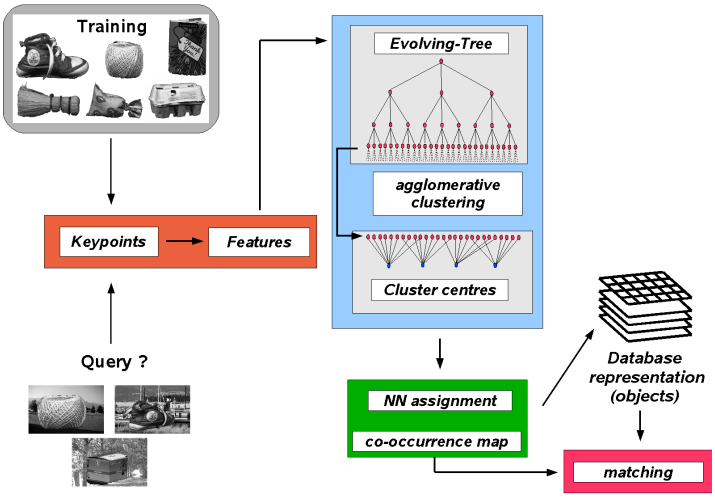


Figure 2: Main components of our co-occurrence recognition system and combination of the Evolving-Tree with the agglomerative clustering algorithm.

2.1 Obtaining the visual vocabulary

To cope with the large number of descriptors obtained by state of the art key-point detectors, we need to cluster them efficiently. Standard clustering methods such as k-means or agglomerative clustering cannot handle such a large amount of descriptors in an efficient manner. Therefore we use a novel combination of two clustering algorithms with different properties. The first one is a tree based variant of a self organizing map, namely the Evolving-Tree proposed by Pakkanen *et al.* [20]. The Evolving-Tree randomly takes training samples from a data-set and uses nearest neighbor assignments until a leaf node is reached. Similar to self organizing maps, the leaf node location is updated with a Kohonen learning rule and once a certain number of leaf node members is reached, the node splits up into a predefined number of child nodes. Due to its simplicity the Evolving-Tree is very fast (we need only about 50 seconds to cluster 200K samples on a Intel Xeon 2.80GHz CPU), but unfortunately the clustering is very imprecise. In accordance with Pakkanen *et al.* [20] we have observed, that the global nearest neighbor property is not fulfilled for most points indexed by the tree. Although the problem can be reduced by pruning the tree, the results remained unsatisfactory. Thus we use a second clustering algorithm and re-cluster the prototypes of the obtained cluster centers in the pruned Evolving-Tree. This second clustering algorithm is a modification of 'agglomerative hierarchical clustering' similar to that one used in [16]. The main advantage of this algorithm is the fact, that we have to select only the tolerated dissimilarity of two points belonging to the same cluster in the feature space. So we can combine the advantages of the two algorithms: (1) the speed of the Evolving-Tree allows us to roughly quantize a feature space even with a very large number of samples in feasible time and (2) the agglomerative clustering algorithm guarantees the necessary similarity of samples assigned to a single cluster. So starting with typically about 200K descriptors we use for creating the visual vocabulary, we end up in a typically size of about $k = 10K$ clusters in a few minutes. The tree also speeds up the search during the recognition as we quickly traverse the Evolving-Tree down to the leaf node and make use of the obtained mapping from agglomerative clustering in the training step.

2.2 Training and recognition of objects, the co-occurrence matrix

Once we have obtained the cluster-centers, training of the database is straight forward. Every object is presented to the recognition system and key-points and descriptors are calculated. Every descriptor is assigned to the best matching cluster center, so that for every interest-point a cluster label is stored. To calculate the co-occurrence matrix, we identify the n nearest neighbors (typically $n = 3$) in image space for every key-point. Thus, every co-occurrence is identified by a pair of cluster centers, which we insert into a two dimensional co-occurrence matrix. The rows and columns of the co-occurrence matrix are the cluster labels of the two key-point descriptors. The co-occurrence matrix is very sparsely populated. Typically only $1 - 2‰$ of the possible co-occurrences are assigned. Thus it is possible, to store only the binary information, whether a co-occurrence is present or not. This binary coding and a sparse storage schema allows us to reduce the necessary amount of memory to a minimum. To build the full representation for a single object (multiple viewpoints), all the co-occurrences of the training images are entered in one single matrix. Therefore, we have exactly one co-occurrence matrix per object trained and the final dimensionality of the training data representation is given by the squared number of cluster centers times the number of objects represented in the database. For the recognition of objects with the co-occurrence matrices we follow the same steps as for training, but for a single query image and obtain a single co-occurrence matrix for each one. The matching procedure itself is deliberately kept very simple. We calculate the matching score for every object representation of the training database by simply AND operation of the co-occurrence matrices and count the number of resulting matches. So in fact, the matching is only a simple maximum voting of congruent co-occurrences in the binary matrices.

3 Experiments

The main purpose of our experiments is to show, that with co-occurrences we can obtain respectable and reliable recognition results even with a very weak and fast descriptor. We do that by calculating ‘recall’-rates for a challenging recognition scenario up to 900 different objects on substantially cluttered background. We first present the database and the reference system we compare our results with, and then several experiments highlighting different aspects of our approach.

3.1 The object recognition task

For our first experiment, we use a subset of the first 50 objects from the publicly available Amsterdam Library of Object Images (ALOI) [11]. Recognition approaches based on local features require a reasonable number of DoG key-points detected on the objects surface. Therefore we sorted the objects of the whole ALOI with respect to their sum of key-points detected, and shifted the upper and lower 50 ones to the end of the database. Thus objects with no or too few key-points as well as those showing an exceptionally high number are not taken into account. To capture enough variances in the appearances, during the training stage we present 12 views (every 30^o) per object to the system. This is similar to the approach of Murase and Nayer [18], but in contrast to them we use a larger amount of objects, can deal with occlusions and work on substantial background clutter - important necessities for a realistic recognition scenario. So in the recognition step, the system has to recognize an object presented in different viewpoint angles and projected to challenging background images (see Figure 3 for some examples). For all of our experiments we use the publicly available binaries of Lowe’s ‘Difference of Gaussian’ (DoG) detector [14] to obtain rather accurate key-points with high repeatability.

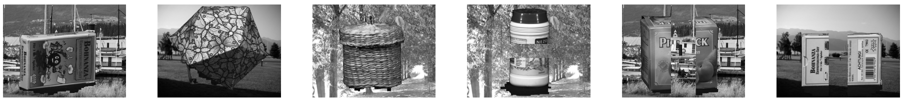


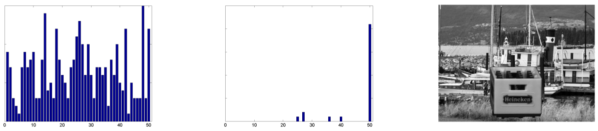
Figure 3: Some examples of objects and partially occluded objects from ALOI database.

3.2 The reference recognition system

To show the impact of the introduced co-occurrence approach, we compare a simple standard voting scheme against our approach. We try to keep the algorithmic parts as much as possible identical for the two systems (Figure 2). The only differences between the reference and co-occurrence recognition system is the representation of the objects (cluster centers versus binary co-occurrence matrix) and the matching method. For the reference system we use an inverse histogram to object voting where every key-point descriptor in a query image is assigned to its nearest neighbor cluster and every cluster votes for one or more objects (similar to [22]). The object with the maximum number of votes is selected. It is essential for the reference recognition system to use only very distinctive feature to cluster assignments, rejecting features near the separation plane of different cluster centers. We follow the approach of Lowe in [14] and use the ratio between the first and second nearest neighbor cluster as a good measure for that property (a distance ratio of $r = 0.8$ gives the best results in our experiments).

3.3 Decreasing information content of the descriptor

The goal of this experiment is to show, that even by decreasing the information content of a descriptor, our approach still keeps feasible recall rates and the matching reliability is higher than for the reference recognition system. As a descriptor we selected PCA-SIFT, a variant of a gradient based descriptor introduced by Ke and Sukthankar [12]. The descriptor is a PCA decomposition of 2 orthogonal gradient images and it gives similar performance rates like SIFT, even if the dimensions are reduced down to 20 [17]. In order to simulate continuously diminishing information content, we can successively remove the less important values of the PCA-projection indicated by the magnitude of the original eigenvalues. We trained it for each dimensionality reduction factor separately as we want to allow the special 'adaptation' of the Evolving-Tree to the desired dimensionality. The main results of the experiment (see Figure 1a) have already been mentioned in the introduction. For decreasing PCA dimension (weaker descriptor) the recognition rates decrease as expected, but the single descriptor approach breaks down much earlier. This is in accordance to the observations of Ke and Sukthankar proposing a dimensionality of $n = 20$ as a good trade-off between matching speed, storage requirements and good recognition results [12]. The co-occurrence approach shows excellent stability and recall performance. It is able to produce nearly perfect recognition results for only 5 or 6 PCA dimensions. Even for 2 PCA dimensions the recall rate is rather high. This experiment indicates, that the co-occurrence approach is not only much more discriminative, but also can deal with 'weaker' descriptors while keeping high recall rates. Furthermore we investigate the significance of the voting processes for the reference and co-occurrence algorithm. Even by manual inspection of the voting histograms, the higher significance of the co-occurrence approach is obvious (Figure 4). In order to get a quantitative estimate of the 'voting significance' we calculate the kurtosis of the voting histogram function. The kurtosis is a statistical measure for the 'peakedness' of a function. The 'ideal kurtosis' (impulse function) for a perfect voting histogram of 50 different objects is $k_{opt} = 48.0204$. In Figure 1b one can see the kurtosis of the reference and our co-occurrence algorithm for a decreasing number of PCA components. The difference of the 'voting significance' measure is obvious even for the highest number of PCA dimensions tested ($n = 36$). So taking into account the results of Figure 1a we can state, that even for comparable recognition rates, the significance of the voting histograms is much higher for the co-occurrence approach resulting in increased reliability of recognition.



(a) (b) (c)

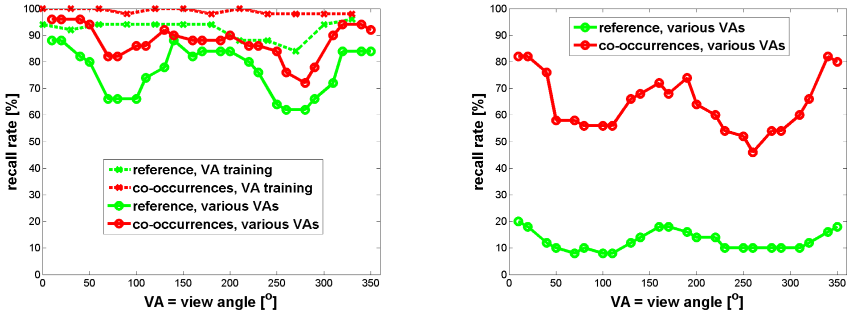
Figure 4: Example of voting histogram for the reference algorithm (a) and our co-occurrence approach (b). The corresponding object is shown in column (c).

3.4 Co-occurrences with SIFT descriptors

In this experiment we show the impact of co-occurrences for standard SIFT-keys as they are a standard descriptor for recognition systems. The recall rates for different viewing angles (VA) can be seen in Figure 5a. We split the recognition results into different curves for the viewing angles originally presented to the system during training (dotted line) and the slightly rotated views of the object on the background image (full line). The recall rate for the co-occurrence approach on 'trained' objects (objects presented to the system in the training stage) is nearly 100%. The improvement obtained by co-occurrences is only about 10% to 15% for various viewpoint angles. Although the improvement introduced through co-occurrences is obvious, the difference is not really large, but SIFT descriptors are already very distinctive features, so that for such a 'strong' descriptor the high performance rates even for a simple voting algorithm are not surprising.

3.5 Spin-Images as an example for a 'weak descriptor'

In this experiment we want to demonstrate the power of our approach applied to a 'real' weak descriptor. A disadvantage of many gradient based descriptors is the necessity to normalize the patches to the principal direction of the region to obtain rotational invariance. Besides the fact, that orientation estimation is not always correct, sometimes more than one orientation is predominant, so that even two or more variants of a patch have to be considered [14]. So for a weak descriptor we prefer a simple, rotational invariant descriptor ideally working on gray values so that we can completely avoid gradient calculation and orientation estimation. A good candidate for such a weak, rotational invariant descriptor is a 2D modification of the original spin-image descriptor proposed by Lazebnik *et al.* [13]. It can be calculated very efficiently and has shown low(!) performance compared to most other descriptors as *e.g.* shown in [17]. Nevertheless in combination with co-occurrences even that poor descriptor can lead to encouraging recognition rates. Figure 5b shows the results of our experiment. As expected, the recall rate for our method consistently outperforms the reference algorithm and an average performance boost of about 60% is obtained. The decrease of performance between 50 to 100 and 250 to 300 degrees is due the fact, that most of the object are captured from the largest side.

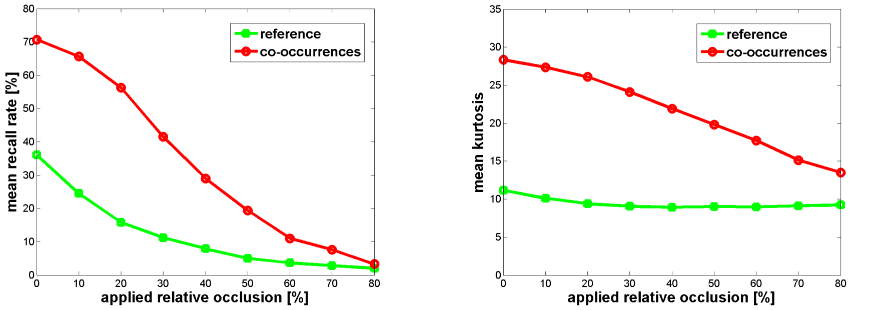


(a) (b)

Figure 5: Recall rates for our co-occurrence algorithm and the reference system on first 50 objects of ALOI database. (a) SIFT descriptor and (b) 'spin images' as a weak descriptor.

3.6 Recognition of partially occluded objects

To support the claim on the improvement of recognition reliability, we repeated the experiments of the last section with the spin image descriptor, but with varying partial occlusions of query images. The occlusions are simulated by transparent rectangles, covering a substantial part of the objects appearance (Figure 3). Transparency means, that we can view the random background on the occluded parts of the objects. In order to avoid biasing the results for elongated objects in certain directions, we have averaged the recall rates for vertical and horizontal occlusions. Figure 6 shows the mean recall rates (a) and kurtosis values of the voting histograms (b) for a different amount of occlusions on the 'sea' background. Besides the already demonstrated fact, that our co-occurrence approach shows consistently better recall rates, one can see the different curvature of the recall function in the range from zero up to 40% recall rate. Thus, the relative decay of recognition rates with respect to the initial values is significantly higher for the standard algorithm those indicating the higher stability of our proposed approach. The same is evident by means of the constantly higher kurtosis values for co-occurrences.

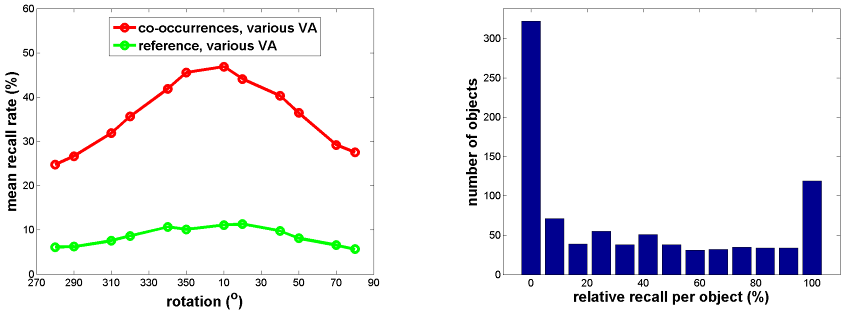


(a) (b)

Figure 6: Mean recall rates (a) and kurtosis values (b) for a different amount of occlusions.

3.7 Scaling behavior and vocabulary generalization

In order to investigate the scaling behavior of our approach, we extended the recognition task to the first 900 objects (6 views every 30^o trained) of the ALOI database. Furthermore we use only the first 100 objects to learn the visual vocabulary which results consistently in about 10K clusters as in the experiments before. Thus, there is no special adaptation of the vocabulary to the whole database (generalization of vocabulary). Figure 7a shows the obtained recall rates for the SIFT descriptors. The difference between the reference system and our proposed algorithm is obvious. The experiment consistently supports the improvement of recall rates and higher matching reliability of the feature co-occurrence algorithm even for much more objects. Nevertheless, despite considering the natural increase of possible mismatches due to the database size, the recall rate is not that high and needs some further consideration. Figure 7b depicts a histogram illustrating the number of objects for a specific relative recall rate. 15% of the objects are perfectly recognized, while about one third of the objects in the database do not work at all. Inspecting those objects we have found the following problems: (1) there are many objects which do not



(a) (b)

Figure 7: (a) Recall rates for our co-occurrence algorithm and the reference system (SIFT). (b) Number of objects with a specific relative recall rate of the first 900 objects.

have a sufficient number of key-points that are detected (*e.g.* no texture, very small objects). (2) some objects can be discriminated only using color, since we use a gray-level based representation these objects are indistinguishable for our method. (3) highlights from the acquisition set-up introduce consistent false matches.

4 Summary and conclusion

In this paper we have demonstrated, that only binary co-occurrences of quantized descriptors are sufficient to build a reliable and efficient object recognition system. Besides the simplicity of the approach the main novelty is the fact, that we do not use the spatial relations for verification of tentative matches or enrichment of other extensive object representations. In our approach we use the co-occurrences alone for object representation and recognition. In order to obtain reliable co-occurrences we present a fast hierarchical quantization algorithm and by limiting the representation of an object to a solely binary representation we can heavily reduce the storage requirements and limit the matching to a very simple and crude algorithm. In a central experiment we could verify, that even by substantially decreasing the information content of a PCA-SIFT descriptor our approach still keeps feasible recall rates and the matching reliability for the feature co-occurrence is much higher, than for a standard recognition approach. Furthermore we have shown, that the approach works well with ‘spin-images’ as an recognition example for a weak but easy computable descriptor, even for substantial occlusion. A final scaling experiment demonstrates the performance increase caused by the novel representation with respect to the bag of key-points approach. The conclusion of the paper is the fact, that instead of further increasing the distinctiveness of certain descriptors and applying computationally expensive matching algorithms, one can also take simple ‘weak’ descriptors and achieve the distinctiveness by the simple concept of binary feature co-occurrence.

References

[1] Ankur Agarwal and Bill Triggs. Hyperfeatures – multilevel local coding for visual recognition. In *Proceedings ECCV*, volume 3951 of *LNCS*, pages 30–43. Springer, 2006.

[2] Guillaume Bouchard and Bill Triggs. Hierarchical part-based visual object categorization. In *Proceedings CVPR*, volume 1, pages 710–715, 2005.

[3] Gustavo Carneiro and David Lowe. Sparse flexible models of local features. In *Proceedings ECCV*, volume 3953 of *LNCS*, pages 29–43, 2006.

[4] David J. Crandall and Daniel P. Huttenlocher. Weakly supervised learning of part-based spatial models for visual object recognition. In *Proceedings ECCV*, pages 16–29. Springer, 2006.

[5] Gabriella Csurka, Christopher R. Dance, Lixin Fan, JuttaWillamowski, and Cedric Bray. Visual categorization with bags of key-points. In *Proceedings SLCV*, 2004.

[6] Jamie Shotton et al. *TextonBoost*: Joint appearance, shape and context modeling for multi-class object recognition and segmentation. In *Proceedings ECCV*, pages 1–15. Springer, 2006.

[7] Josef Sivic et al. Discovering objects and their location in images. In *Proceedings ICCV*, volume 1, pages 370–377, 2005.

[8] Mark Everingham et al. The 2005 pascal visual object classes challenge. In *Selected Proceedings of the First PASCAL Challenges Workshop (LNAI), Springer-Verlag*, April 2006.

[9] Rob Fergus, Pietro Perona, and Andrew Zisserman. Object class recognition by unsupervised scale-invariant learning. In *Proceedings CVPR*, volume 2, pages 264–271, June 2003.

[10] Rob Fergus, Pietro Perona, and Andrew Zisserman. A sparse object category model for efficient learning and exhaustive recognition. In *Proceedings CVPR*, 2005.

[11] Jan Mark Geusebroek, Gertjan Burghouts, and Arnold Smeulders. The Amsterdam library of object images. *International Journal of Computer Vision*, 61(1):103–112, January 2005.

[12] Yan Ke and Rahul Sukthankar. Pca-sift: A more distinctive representation for local image descriptors. In *Proceedings CVPR*, volume 2, pages 506–513, 2004.

[13] Svetlana Lazebnik, Cordelia Schmid, and Jean Ponce. Affine-invariant local descriptors and neighborhood statistics fortexture recognition. In *Proceedings ICCV*, pages 649–655, 2003.

[14] David Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60:91–110, 2004.

[15] Jiri Matas, Ondrei Chum, U. Martin, and Tomas Pajdla. Robust wide baseline stereo from maximally stable extremal regions. In *Proceedings BMVC*, volume 1, pages 384–393, 2002.

[16] Krystian Mikolajczyk, Bastian Leibe, and Bernt Schiele. Local features for object class recognition. In *Proceedings ICCV*, pages 1792–1799. IEEE Computer Society, October 2005.

[17] Krystian Mikolajczyk and Cordelia Schmid. A performance evaluation of local descriptors. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(10):1615–1630, 2005.

[18] Hiroshi Murase and Shree K. Nayar. Visual learning and recognition of 3d objects from appearance. *International Journal of Computer Vision*, 14(1):5–24, Januar 1995.

[19] Andreas Opelt, Axel Pinz, and Andrew Zisserman. Incremental learning of object detectors using a visual alphabet. In *Proceedings CVPR*, volume 1, pages 3–10, 2006.

[20] Jussi Pakkanen, Jukka Iivarinen, and Erkki Oja. The evolving tree-analysis and applications. *IEEE Transactions on Neural Networks*, 17(3):591–603, 2006.

[21] Cordelia Schmid and Roger Mohr. Local grayvalue invariants for image retrieval. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19:530–535, 1997.

[22] Josef Sivic and Andrew Zisserman. Video Google: A text retrieval approach to object matching invideos. In *Proceedings ICCV*, October 2003.

[23] Hongming Zhang, Wen Gao, Xilin Chen, and Debin Zhao. Object detection using spatial histogram features. *Image and Vision Computing*, 24(4):327–341, April 2006.