Object Detection Using Multi-Local Feature Manifolds

Oscar Danielsson, Stefan Carlsson, Josephine Sullivan
Royal Inst. of Technology, Stockholm, Sweden
{osda02 | stefanc | sullivan}@csc.kth.se

Abstract

Many object categories are better characterized by the shape of their contour than by local appearance properties like texture or color. Multi-local features are designed in order to capture the global discriminative structure of an object while at the same time avoiding the drawbacks with traditional global descriptors such as sensitivity to irrelevant image properties. The specific structure of multi-local features allows us to generate new feature exemplars by linear combinations which effectively increases the set of stored training exemplars. We demonstrate that a multi-local feature is a good "weak detector" of shape-based object categories and that it can accurately estimate the bounding box of objects in an image. Using just a single multi-local feature descriptor we obtain detection results comparable to those of more complex and elaborate systems. It is our opinion that multi-local features have a great potential as generic object descriptors with very interesting possibilities of feature sharing within and between classes.

1. Introduction

The use of local feature descriptors has been a major successful tool for object classification in recent years. Local descriptors can be designed in order to avoid the disturbing effects of class irrelevant foreground and background clutter in the image. Their limited spatial extent means that intra class variation among spatially corresponding features is limited, making them efficient for classification. They have been especially successful in the detection of object classes such as faces, cars, motorcycles [3, 11, 14, 16, 20, 21] where well defined subparts such as eyes, nose, wheels etc. can be captured as local features. For a large set of object classes however, well defined subparts are difficult to define and extract. Even if they can be defined, their intra class variability is sometimes very large, or they are not very discriminative [7, 18]. These classes in general need to be characterized by their global shape properties. This typically applies to a large range of "simple" objects like cups, bottles and tools etc.

The problem of extracting and representing global object structure in an efficient way has a very long history going back to theories of perceptual grouping and gestalt perception. Just simply increasing the spatial size of the local descriptors will in general have the result that they become susceptible to the effects of class irrelevant clutter in the image [15, 19], although recent modifications based on subdivision and boosting have proved to be comparatively efficient [8, 9]. In order to avoid this, grouping local structure therefore in general takes place along the boundaries and edges of an object, see e.g. the recent work in [4, 5]. The problem of grouping is however very difficult since it requires the extraction of object shape by edge detection or some other process which is often unreliable. In general one has to be content with the extraction of short fragments of object shape that may not have sufficient discriminatory power.

What is needed is a descriptor that preserves all the good clutter rejection properties of local descriptors while at the same time being able to capture global shape. This seems like a contradiction but does not necessarily have to be so. We will demonstrate that it is possible to define efficient *multi-local* features which are just specific spatial constellations of local features that actually have these desirable properties. In addition to this, exemplars of multi-local features extracted from training images can be used to generate linear manifolds of multi-local feature classes that effectively increase the training data set of features. Having a generative model for a feature class like this is a very powerful tool for classification. Classification procedures can be based on manifolds instead of individual exemplars.

2 Multi-local features

A multi-local feature consists of a set of k local features at locations $(\mathbf{x_1} \dots \mathbf{x_k})$ with local content characterized by descriptors $(c_1 \dots c_k)$. Local content can be any descriptor but in this work it will typically consist of just directionality information. Local descriptors can therefore be applied to any part of the object's shape outline. By selecting corre-

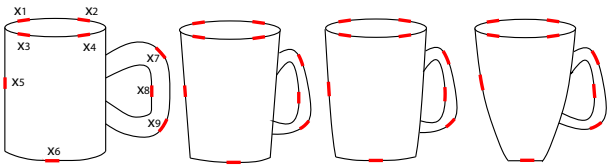


Figure 1. *Multi-local features consist of specific spatial constellations of local features. Corresponding locations on examples from an object class then generate a specific multi-local feature manifold for that class*

sponding locations for local features across examples of an object class, we end up with a set of exemplar descriptors $(\mathbf{x_1^{(i)}}, c_1^{(i)} \dots \mathbf{x_k^{(i)}}, c_k^{(i)})$ for training exemplars $i = 1 \dots n$ for a specific feature.

The multi-local feature class manifold will consist of all possible values of the descriptor: $(\mathbf{x_1}, c_1 \dots \mathbf{x_k}, c_k)$ that are extracted from corresponding locations $\mathbf{x_i}$ in the image exemplars of the object class. It will include all possible image translations and scalings of the individual examples. Typically it will depict a specific viewpoint of a class as in the standard constellation models considered in [3, 21].

If we consider simple local structure descriptors (c_i) such as just directionality, the multi-local feature will capture global shape properties of the object class. This is in contrast to standard appearance based global image descriptors that will have to deal with a large range of appearance variation that is irrelevant to the object class. The prize we pay for this interesting property is of course an increased complexity in the extraction stage.

Multi-local features can have a range of variation of their number of local feature components and their spatial extent. By considering a small spatial extent, the multi-local feature descriptor will be just an enlarged rich local descriptor such as that considered in e.g. [4, 5]. The number of component local features in a multi-local feature will clearly affect the manifold of features occurring in images. If the number is small, we will get very general features that can be expected to be shared across object classes while if the number and spatial extent is large, we get very specific features that will serve more as shape templates.

The size and complexity of the multi-local feature obviously affects the complexity of extracting it and it's discriminative value when used for e.g. object detection. In this paper we will focus on these issues of complexity and performance and demonstrate the applicability of the concept of multi-local features. In future work we will consider the broader issues of integration of several multi-local features, feature sharing among object classes and automatic discovery of multi-local features.

The idea of multi-local feature is closely related to the concept of constellations [3] and pictorial structures [2] although these in general assume rich local descriptors. In our multi-local descriptor, the local features can be very simple and the information of the feature is contained in the spatial arrangement rather than local content. This idea of looking at spatial arrangements and relations of simple features occurs in [1, 19] and recently also in [10].

2.1 Multi-local feature detection

For detection we are faced with the problem of deciding whether the image contains a multi-local feature with a descriptor in the class manifold. Even for values of k in the range $5 - 10$ this can be a huge search problem. In order to solve this we will exploit the specific structure and properties of the multi-local feature manifold. Ideally we should be able to characterize the manifold by a complete statistical description $p(\mathbf{x_1}, c_1 \dots \mathbf{x_k}, c_k)$. We will estimate this from labeled exemplars of images of the class. This will allow us to define conditionals such as $p(\mathbf{x_j}, c_j | \mathbf{x_1}, c_1 \dots \mathbf{x_{j-1}}, c_{j-1})$ which will be used to detect local features efficiently in a recursive manner.

The spatial constellations of multi-local features for a specific class and viewpoint will depict the internal class shape variation. In general this variation is smooth in the sense that similar exemplars can be deformed into each other. Any intermediate deformation will then represent a valid exemplar of the class. This is a purely empirical observation and has been validated experimentally on the data sets that we have used so far. The important thing is that this allows us to vastly extend the set of exemplars from those sampled from training images into the whole convex hull:

$$C = \left\{ (\mathbf{x_1} \dots \mathbf{x_k}) \mid \begin{array}{l} \mathbf{x_j} = \sum_i \alpha_i \mathbf{x_j^{(i)}},\ j = 1 \dots k \\ \sum_i \alpha_i \ = \ 1 \end{array} \right\} \quad (1)$$

This is illustrated in figure 2 showing extracted multi-local features from bottle exemplars and their interpolations. This property will be exploited in a final verification stage for hypothesized multi-local features in an image.

The detection of multi-local features in an image will proceed by a hypothesis generation and verification process. From a set of n labeled training images containing exemplars of an object class we manually extract multi-local features: $(\mathbf{x_1^{(i)}}, c_1^{(i)} \dots \mathbf{x_k^{(i)}}, c_k^{(i)})$, $i = 1 \dots n$. In this work we will only consider local feature properties c_j depicting the directional structure of the image location $\mathbf{x_j}$. We will initiate the search for multi-local features in a test image by first computing a binary feature map indicating the presence or not of the various local features contained in the multi-local constellation. The multi-local training features are then used to generate hypotheses of the existence

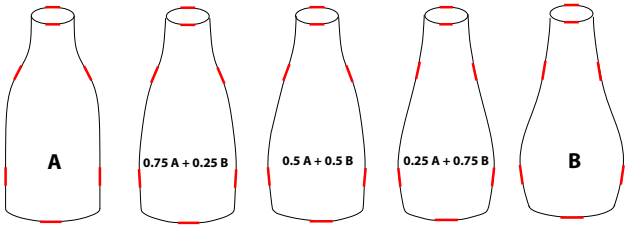


Figure 2. *Convex linear combinations of multi-local features tend to generate multi-local features that emanate from valid representative exemplars in the class. The three middle bottle multi-local features above are generated by linear combinations of the multi-local features of the left-most* **A** *and right-most* **B** *bottles with linear weights successively changing*

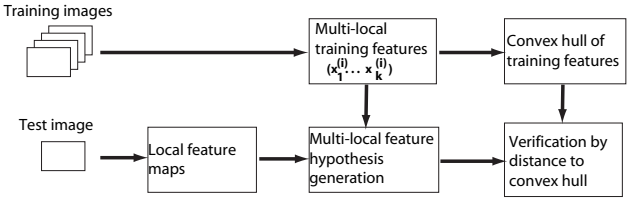


Figure 3. *Overview of multi-local feature hypothesis and verification process using estimated distribution from training images. Locations* $(\mathbf{x_1} \ldots \mathbf{x_k})$ *of components in a multi-local feature are hypothesized using the set of labeled training exemplars and verified by computing their distance to the convex hull of the training feature locations*

of multi-local features in the test image. These are subsequently verified by computing their distance to the convex hull of the multi-local training features as given by eq. 1. The whole procedure is illustrated in figure 3.

An example of the algorithm is shown in figure 4. Here we search for a multi-local feature consisting of four local features in a box-like configuration. The first local feature (defined by c_1) is present at a particular location if there is a sufficiently strong gradient with direction sufficiently close to horizontal at that location (a). For each occurrence of the first local feature, we search for plausible occurrences of the second local feature using the conditional $p(\mathbf{x_2}|\mathbf{x_1})$, which is defined up to an unknown scale factor. This yields a conic region in the image, which is searched for the second local feature (b). The second local feature is defined by c_2 to be a sufficiently strong gradient with direction close to vertical. If the second local feature is found we finally sample from $p(\mathbf{x_4}, \mathbf{x_3}|\mathbf{x_2}, \mathbf{x_1})$ to get plausible locations of the last two local features. We use the Euclidean (or Chamfer)

Figure 4. *Illustration of the algorithm for finding instances of a given multi-local feature. For each occurrence of the first local feature (a), we search for occurrences of the second local feature (b). We then sample plausible locations of the other local features using the training exemplars and check for presence of these local features at the predicted positions (c). If all local features are present, we sample their constellation and compute the distance to the closest point on the convex hull of training exemplars*

distance transforms of the binary feature maps of these two local features to determine if they are present in the image sufficiently close to the predicted positions (c). If so, we sample the image locations of the local features and compute the distance to the convex hull of multi-local features extracted during training. The final output of the algorithm is the image locations of the local features and the distance to this convex hull for each detection. A detailed description of the algorithm is given in figure 5.

The computational complexity of multi-local feature detection is dominated by the hypothesis generation stage. The complexity of hypothesis generation grows proportionally to the number of occurrences of the first local feature, c_1, times the number of occurrences of the second local feature, c_2, times the number of samples drawn from $p(\mathbf{x_4}, \mathbf{x_3}|\mathbf{x_2}, \mathbf{x_1})$.

3 Experiments

The experiments presented in this section aim to illustrate the following properties of multi-local features:

1. A single multi-local feature is a good weak detector and can produce accurate estimates of the bounding box of instances of the target object class with relatively few false positives.

2. The detection performance does not necessarily improve by adding more local features to the multi-local feature.

We have experimented with three shape-based object classes: apple logos, bottles and mugs. For each class, we

Inputs: image I, exemplars $(\mathbf{x_1^{(i)}}, c_1 \ldots \mathbf{x_k^{(i)}}, c_k)$, $i = 1 \ldots n$, threshold ρ.

1. Compute a binary feature map for each local feature.
2. Compute the Euclidean (or Chamfer) distance transforms on these feature maps.
3. For each occurrence $\mathbf{x_1}$ of the first local feature c_1:
   - For each occurrence $\mathbf{x_2}$ of the second local feature c_2 such that $p(\mathbf{x_2}|\mathbf{x_1}) > \rho$:
     - Sample from $p(\mathbf{x_k}, \ldots, \mathbf{x_3}|\mathbf{x_2}, \mathbf{x_1})$
     - If local features $c_k, \ldots, c_3$ are present close to the hypothesized locations $\mathbf{x_k}, \ldots, \mathbf{x_3}$
       * Save image locations of $c_k, \ldots, c_1$
4. For each multi-local feature $(\mathbf{x_1}, c_1, \ldots, \mathbf{x_k}, c_k)$ sampled from the image, compute the distance to the convex hull of training exemplars by solving the following quadratic program:

$$d = \min_{\mathbf{w}\in\Re^n, \mathbf{t}\in\Re^2, \mathbf{r_j}\in\Re^2} \frac{1}{2}\sum_{j=1}^{k} \|\mathbf{r_j}\|^2 \quad (2)$$

subject to

$$\sum_{i=1}^{n} w_i \cdot \mathbf{x_j^{(i)}} + \mathbf{t} + \mathbf{r_j} = \mathbf{x_j}, j = 1 \ldots k$$

$$\mathbf{w} \geq \mathbf{0}$$

5. Normalize the distance by the square of the scale of the multi-local feature in the image.

Output: A list of multi-local feature occurrences in the image along with distances to the convex hull.

Figure 5. *Detailed description of algorithm for finding instances of a given multi-local feature*

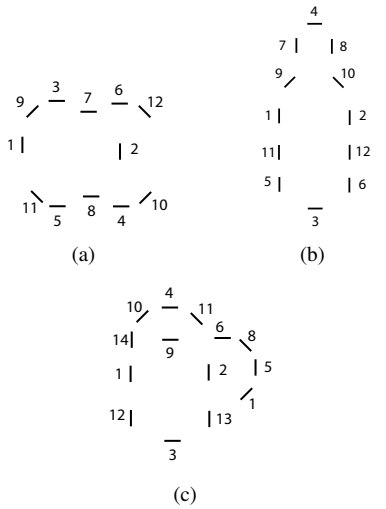


Figure 6. *Typical instances of the large multi-local features selected for (a) apple logos, (b) bottles and (c) mugs. Subsets of these with various numbers of local features are used in the experiments*

have selected a large multi-local feature as illustrated in figure 6. Exemplars of these features were manually clicked in a number of training images downloaded from Google images. We then use sub-features with increasing numbers of local features to perform object detection. We investigate how the detection performance varies with the number of local features.

3.1 Training

In the current experiments, we have manually extracted exemplars $\left(\mathbf{x_1^{(i)}}, c_1, \cdots, \mathbf{x_k^{(i)}}, c_k\right)$ of multi-local features by clicking locations in a number of training images downloaded from Google images. We have used the heuristics presented below to select good multi-local features (in future work, we will evaluate automatic learning methods based partly on these heuristics):

1. Since we are going to use the multi-local features as weak object detectors, we want them to be a good predictor of the bounding box of the object. In general, multi-local features that "span" the object are better in this respect and we try to select multi-local features that capture global rather than rich local shape properties of the object.

2. We try to select multi-local features that describe shape properties that are characteristic of the object but occur with as little variation as possible over the

Table 1. Detection rates of the best multi-local feature at 0.4 FPPI.

Applelogos	Bottles	Mugs
77.3	72.73	60.6

class.

3. Increasing the number of local features makes the multi-local feature more specific and thus reduces the number of false positives. However, including too many local features will result in a multi-local feature that is not shared by all members of the class. Iteratively increasing the number of local features is a good way to find the best trade-off.

3.2 Evaluation

We present an evaluation of the detection performance on the ETHZ Shape Classes dataset [4]. This dataset is challenging due to large intra-class variation, clutter and varying scales. However, object instances are generally not occluded. The whole dataset was used for testing and all images not in the target object category was used as negative examples.

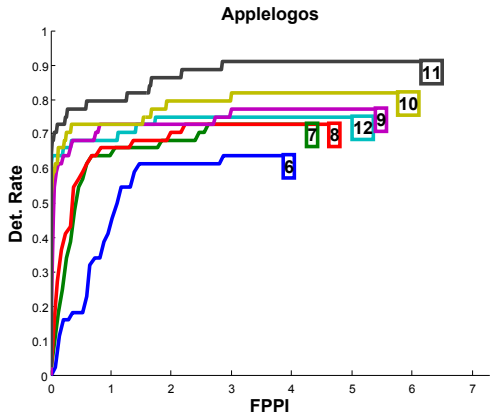
Previous experiments on this dataset have used a very lenient evaluation criterion and count a detection as correct if the bounding box overlaps more than 20 % with the ground truth bounding box and vice versa [4, 5]. In order to get a better measure of the accuracy of the predicted bounding boxes, we use a 70 % overlap criterion instead. In figure 7 the detection rate (Det. Rate) is plotted against the number of false positives per image (FPPI).

4 Results

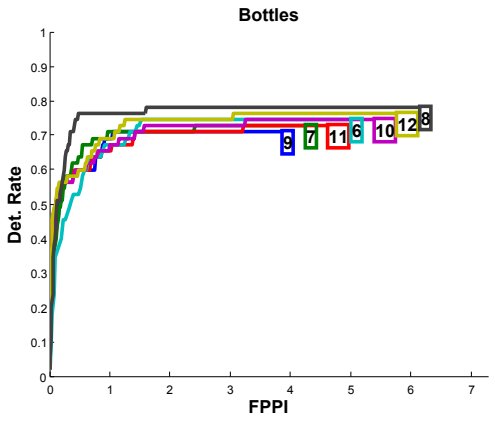
The results indicate that even a single multi-local feature can achieve a quite good detection rate at a reasonable number of false positives per image. In particular for mugs and apple logos we also see the expected effect that including too many local features will reduce the detection rate considerably, since the multi-local feature gets too specific and is no longer shared across the class. The optimal number of local features is 11 for apple logos and 9 for mugs.

In figure 8 we show detections in a few example images from the test dataset. The predicted bounding boxes are marked in green. The logarithm of the matching distance of each detection is given in the upper left corner of the bounding box.

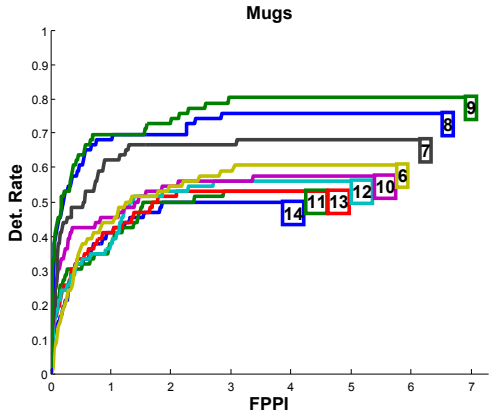
In table 1 we give the detection performance of the best single multi-local feature at 0.4 FPPI. This allows for com-



(a)

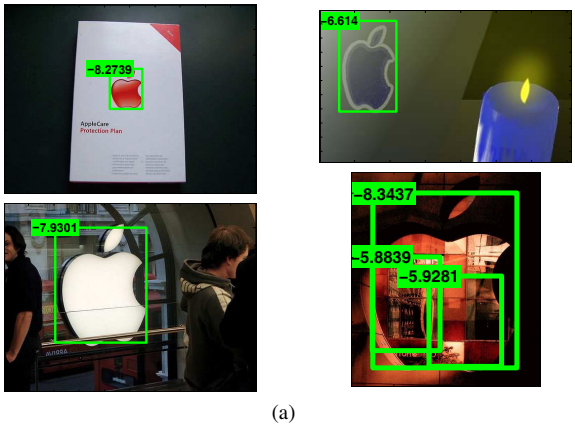


(b)

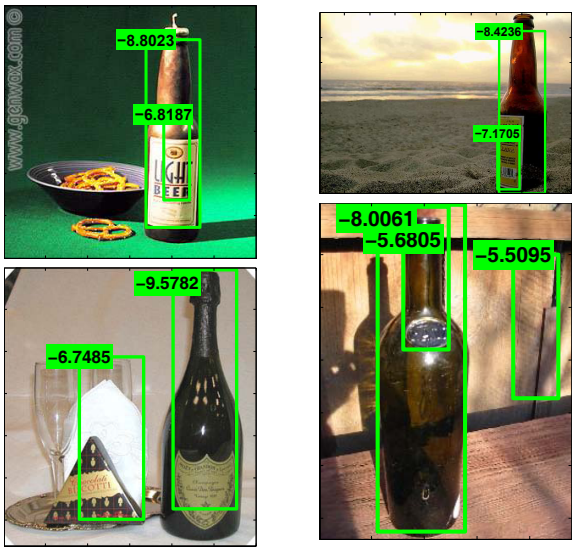


(c)

Figure 7. *The detection rate plotted against the number of false positives per image of (a) apple logos, (b) bottles and (c) mugs. For each object class several curves are plotted, representing different numbers of local features. The number of local features used is given at the end of the curve. A detection is counted as correct if the bounding box overlaps more than 70 % with the ground truth and vice versa*



(a)



(b)

(c)

Figure 8. *Example detections of (a) apple logos (using 12 local features), (b) bottles (using 9 local features) and (c) mugs (using 8 local features). At the upper left corner of each bounding box is the logarithm of the detection distance (lower is better). Detections with distances better than -5.5 are shown*

parison to Ferrari et. al. [5], but we stress that we are using a much stricter evaluation criterion (they count a detection as correct if the bounding box overlaps more than 20 % with the ground truth bounding box and vice versa, while we require 70 % overlap). Ferrari et. al. present detection rates at 0.4 FPPI on the ETHZ Shape Classes dataset using a Hough-voting scheme and using their full shape matching system. Using only Hough-voting they get 35.9 %, 71.7 % and 51.4 % detection rates for apple logos, bottles and mugs respectively. Using their full system they get 83.2 % 83.2 % 83.6 %. Comparing that to the values in table 1, we see that even a single multi-local feature gives comparatively good detection performance on these object categories.

5 Discussion and Conclusion

We have investigated the use of linear manifolds of constellations of simple local features (multi-local features) for object detection. If these constellations "span" the bounding box of the object, this description captures global shape properties of the object. We have claimed that

- A multi-local feature is a good weak object detector and even a single multi-local feature might yield a detection performance comparable to a full object detection system.
- All convex combinations of a given set of valid multi-local feature exemplars are also reasonable exemplars of the target shape. We thus have a good generative model and the detection procedure is based on generating and verifying hypotheses. Feature detection can therefore be based on *manifolds* generated by training exemplars instead of the training exemplars themselves.
- Since our method uses a sparse set of local feature occurrences, it is less dependent on the quality of the edge detector than methods using locally connected edge segments [4, 5, 6, 17, 13]. This robustness of multi-local features will be very relevant when we consider object detection in severe occlusion situations.

However, the detection of multi-local features in its current implementation is sensitive to occlusions and can only represent unimodal shape variations. Therefore, combinations of several multi-local features are necessary to achieve a flexible object detector. The optimal selection and combination of multi-local features for recognition will be a topic of future studies. The problem will be to define sets of multi-local features for a specific class with optimal discrimination capabilities. In view of the results, it seems reasonable to investigate methods that sequentially increase the number of local features until further additions

no longer improve detection performance. A natural extension to building sequences of multi-local features of increasing size is to build hierarchies of multi-local features, where the multi-local features represented by the leaf nodes share common sub-features.

References

[1] Carlsson S., (1999) Order Structure, Correspondence and Shape Based categories International Workshop on Shape, Contour and Grouping may, Springer Lecture Notes in Computer Science 1681

[2] Felzenszwalb, P.F., Huttenlocher, D.P. Pictorial Structures for Object Recognition, IJCV(61), No. 1, January 2005, pp. 55-79.

[3] Fergus, R., Perona, P., Zisserman, A., Object class recognition by unsupervised scale-invariant learning, Proc. CVPR03(II: 264-271).

[4] Ferrari, V., Tuytelaars, T. and Van Gool, L., Object Detection with Contour Segment Networks, Proc. of the European Conference on Computer Vision (ECCV), 2006

[5] Ferrari, V., Jurie F. and Schmid, C., Accurate Object Detection with Deformable Shape Models Learnt from Images, Proc. of the IEEE Computer Vision and Pattern Recognition (CVPR), 2007

[6] Jurie, F. and Schmid, C., Scale-Invariant Shape Features for Recognition of Object Categories, Proc. of the IEEE Computer Vision and Pattern Recognition (CVPR), 2004

[7] Jurie, F., Triggs, B., Creating Efficient Codebooks for Visual Recognition, ICCV05(I: 604-610).

[8] Laptev, I., Improvements of Object Detection Using Boosted Histograms, BMVC06(III:949).

[9] Lazebnik, S., Schmid, C., Ponce, J., Beyond Bags of Features: Spatial Pyramid Matching for Recognizing Natural Scene Categories, CVPR06(II: 2169-2178).

[10] Leordeanu M., Hebert M., and Sukthankar R. Beyond Local Appearance: Category Recognition from Pairwise Interactions of Simple Features Proc. CVPR07 (1 - 8)

[11] Lowe D.G. , Object Recognition from Local Scale-Invariant Features, Proc. Intl Conf. on Computer Vision (ICCV99), Corfu, pp. 1150-1157, 1999.

[12] Murphy, K., Torralba, A. and Freeman, W.T.F. Using the forest to see the trees: a graphical model relating features, objects and scenes. In Advances in Neural Information Processing Systems 16 (NIPS 2003), 2004.

[13] Opelt, A., Pinz, A. and Zisserman, A., A Boundary-Fragment Model for Object Detection, Proc. of the European Conference of Computer Vision (ECCV), 2006

[14] Perronnin, F., Dance, C., Csurka, G., Bressan, M., Adapted Vocabularies for Generic Visual Categorization, ECCV06(IV: 464-475).

[15] Schiele, B., Crowley, J.L., Recognition without Correspondence using Multidimensional Receptive Field Histograms, IJCV(36), No. 1, January 2000, pp. 31-50.

[16] Schneiderman, H., A Statistical Approach to 3D Object Detection Applied to Faces and Cars, Proc. CVPR 2000.

[17] Shotton, J., Blake, A. and Cipolla, R., Contour-Based Learning for Object Detection, Proc. of the International Conference of Computer Vision (ICCV), 2005

[18] Stark, M., Schiele, B., How Good are Local Features for Classes of Geometric Objects ?, ICCV07(1-8).

[19] Thuresson J and Carlsson S. (2004) Appearance Based Qualitative Image Description for Object Class Recognition In Proc. 8th European Conf. on Computer Vision (ECCV)

[20] Ullman, S., Vidal-Naquet, M., Sali, E., Visual features of intermediate complexity and their use in classification, Nature Neuroscience(5), No. 7, 2002, pp. 682-687.

[21] Weber, M., Welling, M., Perona, P., Unsupervised Learning of Models for Visual Object Class Recognition, Proc. ECCV 2000, pp 18 - 32 Springer LNCS 1842