Appearance Based Qualitative Image Description for Object Class Recognition

Johan Thureson and Stefan Carlsson

Numerical Analysis and Computing Science, Royal Institute of Technology, (KTH),
S-100 44 Stockholm, Sweden, {johant,stefanc}@nada.kth.se,
`http://www.nada.kth.se/~stefanc`

Abstract. The problem of recognizing classes of objects as opposed to special instances requires methods of comparing images that capture the variation within the class while they discriminate against objects outside the class. We present a simple method for image description based on histograms of qualitative shape indexes computed from the combination of triplets of sampled locations and gradient directions in the image. We demonstrate that this method indeed is able to capture variation within classes of objects and we apply it to the problem of recognizing four different categories from a large database. Using our descriptor on the whole image, containing varying degrees of background clutter, we obtain results for two of the objects that are superior to the best results published so far for this database. By cropping images manually we demonstrate that our method has a potential to handle also the other objects when supplied with an algorithm for searching the image. We argue that our method, based on qualitative image properties, capture the large range of variation that is typically encountered within an object class. This means that our method can be used on substantially larger patches of images than existing methods based on simpler criteria for evaluating image similarity.

Keywords: object recognition, shape, appearance

1   Introduction

Recognizing object classes and categories as opposed to specific instances, introduces the extra problem of within-class variation that will affect the appearance of the image. This is added to the standard problems of viewpoint variation, illumination etc. that induces variations in the image. The challenge is then to devise methods of assessing image similarity that capture these extra within-class variations while at the same time discriminate against objects of different classes. This problem has been attacked and produced interesting results, using quite standard methods of image representation relying heavily on advanced methods of learning and classification [2,3,6,8,9] These approaches are in general all based on the extraction of image information from a window covering part of the object, or in other words, representing only a fragment of the object. The size of

the fragments then effectively controls the within-class variation as different instances within a class are imaged. For a certain method of image representation and similarity assessment, there is in general an optimal size of the fragments in terms of discriminability, [2]. Increasing the size of the fragments beyond this size will decrease performance due to increased within-class variation that is not captured by the specific similarity criterion used. The fragment based approach can also be motivated from the fact that objects should be recognizable also in cases of occlusions, i.e. only a part of the object is available for recognition. It can therefore easily be motivated as a general approach to recognition. However, the main limitation of fragment based approaches still seems to lie in the fact that within class variation is not captured sufficiently by the similarity measures used. Ideally if we could use any fragment size, from very small up to fragments covering the whole object, we would improve the performance of algorithms for object class recognition.

Given any method to asses similarity of images, it's usefulness for *object class recognition* lies in it's ability to differentiate between classes. If similarity is measured normalized between 0 and 1.0 the ideal similarity measure would return 1.0 for objects in the same class and 0 for objects in different classes. In practise we have to be content with a gradually descending measure of similarity as images are gradually deformed. The important thing is that a sufficiently high value of similarity can be obtained as long as the two images represent the same object class.

2   Appearance vs. Shape Variation

Object shape is generally considered as a strong discriminating feature for object class recognition. By registering the appearance of an object in a window, object shape is only indirectly measured. Traditionally, object shape has been looked for in the gray value edges of the image, in general considered to coincide with 3D edges of the object. This however neglects the effects of object shape on the smoothly varying parts of the image which is ideally capture by appearance based methods. The split between edge based methods trying to capture object shape directly and appearance based methods giving an indirect indication of shape is unfortunate since they lead to quite different types of processing for recognition. Very little attention has been paid to the possibility of a unified representation of shape, accommodating both the direct shape induced variations of image edges and the indirect one's in the appearance of the image gray values. We will therefore investigate the use of qualitative statistical descriptors based on combinations of gradient directions in an image patch. The descriptors are based on the order type which encodes the qualitative relations between combinations of multiple locations and directions of gradients as described in [4], [5], [11] where it was used for correspondence computation in sparse edge images. The *histogram of order types* of an image patch is used as a descriptor for the patch. This descriptor then encodes statistical information about qualitative image structure

Fig. 1. The gradients of an image indirectly encode the shape of an object

and therefore potentially captures qualitative variations of images as displayed between the members of an object class
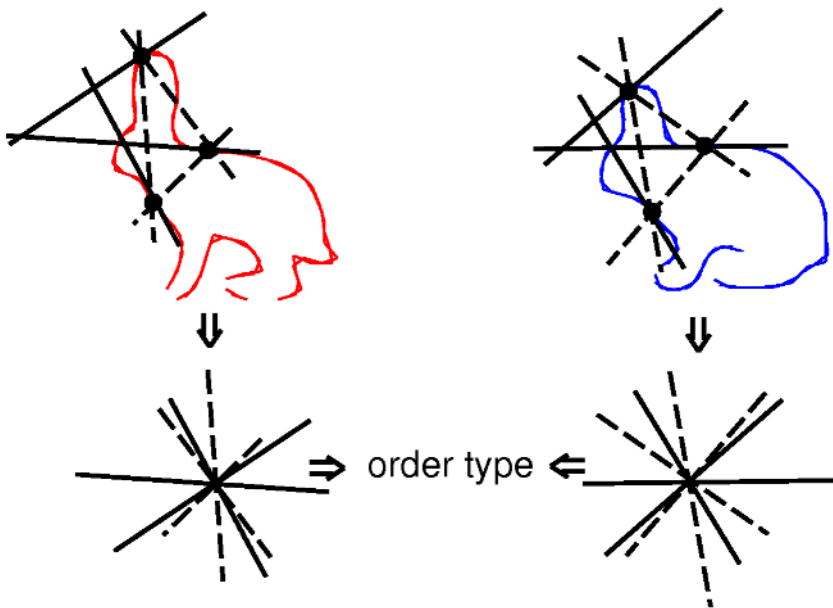

Fig. 2. Three points and their corresponding (orthogonal) gradient directions generates six lines. The angular order of these six lines defines the *order type* of the point-line collection. Corresponding points in qualitatively similar images in general generates the same order type.

If we take three points in an image together with the lines of gradient directions, (strictly speaking the directions orthogonal to the gradients), we can speak about the combinatorial structure of this combined set of points and lines. The three lines have an internal order by considering the angular orientations and the lines are ordered w.r.t. the points by considering the orientation of the lines relative to the points. The idea of *order type* of the combined set can be

easily captured by considering the three lines formed by connecting every pair of points together with the three gradient lines. The *angular order*, of these six lines then defines an index which defines the *order type* of the set. From fig. 2 we see that the order types for perceptually corresponding points in two similar shapes in general are equivalent. The unique assignment of points requires a canonical numbering of the three points for which we use the lowest point as number one and count clockwise. The order type is therefore an interesting candidate for a qualitative descriptor that stays invariant as long as a shape is deformed into a a perceptually similar shape. Order types as defined above are only strictly invariant to affine deformations. For more general types of deformations, invariance of the order type depends on the relative location of the three points. This means that if we consider the collection of all order types that can be computed by considering every triplet of points in an image that has a well defined gradient direction, we get a representation of the entire shape that has interesting invariance properties w.r.t. to smooth deformations that do not alter the shape too drastically. This is often the case between pairs of instances in an object category as noted very early by d'Arcy Thompson [10].
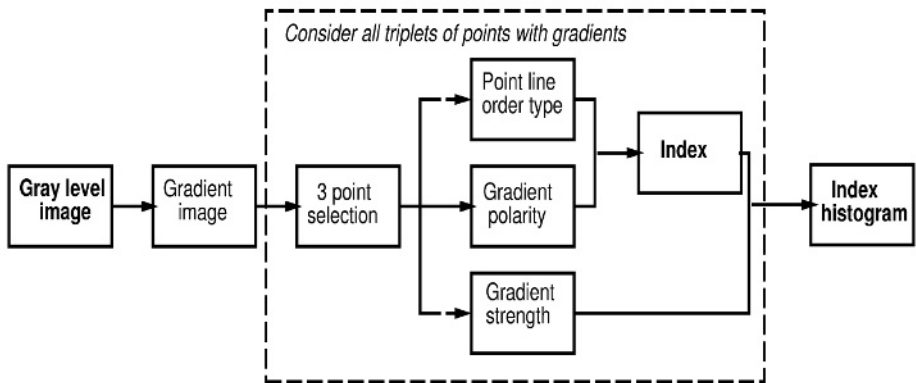

Fig. 3. Procedure for computing *weighted index histogram* from gray level image: A gradient detector is applied to the image. The gradients are thresholded and subsampled. All triplets of remaining gradients are selected and for each triplet we compute an index representing the joint qualitative structure of the three gradients based on order type and polarity. The index is given a weight depending on the strength of the gradients and a weighted histogram of index-occurrences is computed

The procedure for computing the *weighted index histogram* is illustrated by the block diagram of fig. 3. and can be summarized in the following points:

1. The image is smoothed with a Gaussian operator, a gradient detector is applied and it's output thresholded to give a preset total number of gradients
2. For *every combination of three gradients* we compute an *order type index* using the following procedure:

    a) We choose the lowest point as number one and number the others in a clockwise ordering. The locations and angular directions of the three gradients are then denoted as $x_1, y_1, d_1, x_2, y_2, d_2, x_3, y_3, d_3$.

    b) We compute the three directions d_{12}, d_{13} and d_{23} of the lines joining point pairs 12, 13, 23 respectively

    c) The six numbers $d_1, d_2, d_3, d_{12}, d_{13}, d_{23}$ are ranked and the the permutation of the directions $1, 2, 3, 12, 13, 23$ is given a number denoted as the *order type index* for the three gradients

    d) The polarity of each gradient is denoted which increases the index with a factor of $2^3 = 8$

    e) The average strength of the three gradients is computed and used as a weight

3. A histogram of occurrence of the various order type + polarity indexes is computed by considering *all combinations of three gradients* in the image. Each index entry is weighted by the average strength of the three gradients used to compute the index

For the specific way of choosing the order type that we have , we get 243 distinct cases. By considering the polarity of the gradients, we get 1944 different indexes. The histogram of occupancy of these indexes as we choose all combinations of three points defines a simple shape descriptor for the image. In comparison, the same kind of histogram including a fourth point, was used in [4], [5], [11]. By fixating the fourth point a "shape context descriptor" [1] was computed and used for correspondence matching between shapes. Recognition was then based on the evaluation of this correspondence field. By using one single shape descriptor histogram for the whole image we get a comparatively less complex algorithm which allows us to consider every gradient in an image instead of just the selected directions after edge detection as in [4], [5], [11]. The histogram can be seen as a specific way to capture higher order statistics of image gradients where previously mainly first order has been used [7].

3   Comparing Images

Given the histogram vector based on qualitative image properties, comparisons between images can now be performed by simple normalized inner products. If the histogram vector captures qualitative image similarity we would expect images of a certain class to be close using this inner product while images of different classes should be further apart. A simple test of this property was made on two sets of images, motor bikes and faces, with about 15 examples in each class. A subset of those examples are shown in fig 4. Fig 5 shows histograms of inner products between all examples in two classes for the cases face-face, motor bike-motor bike and face-motor bike. This figure clearly demonstrates that pairs of images picked from the same class are definitively closer than pairs picked from different classes. The histograms vectors used are therefore able to capture similarity within a class and at the same time discriminate between different classes.
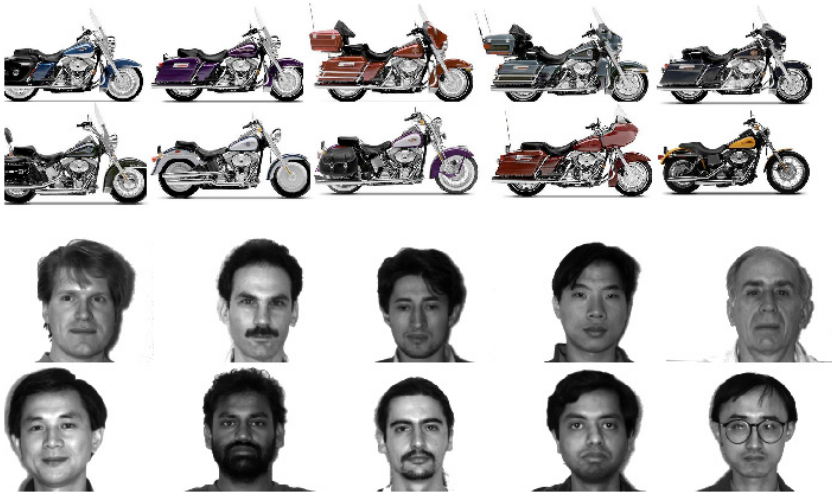
Fig. 4. Subset of images from two classes used for testing similarity measure
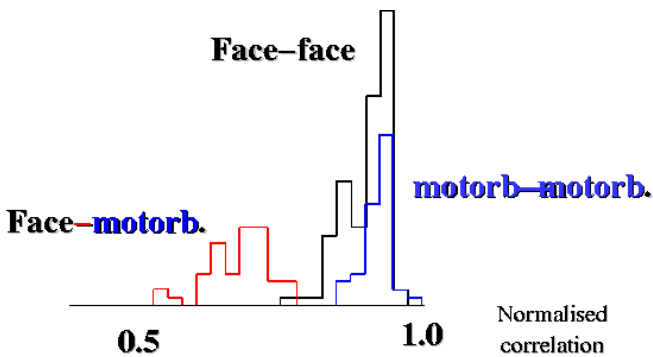

Fig. 5. Histograms of inner products of pairs of histogram vectors.

In another test we computed all inner products between the simple silhouette images of rabbits. Being silhouettes, the image differences are accounted for by shape differences. Using the inner products as a distance measure, we mapped the images in the rabbit set to positions in the plane such that the distance between two mapped rabbit examples corresponds to the distance given by the inner products as well as possible. This procedure is known as multi-dimensional scaling and from this it can be seen that the inner product distance corresponds quite well to a distance measure given by visual inspection of the rabbit examples. Note that there is a gradual transition of the shapes as we move from the upper to the lower part of this diagram
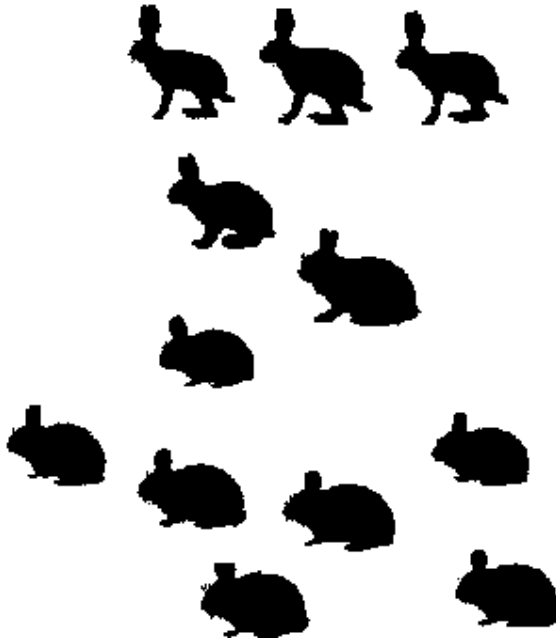
Fig. 6. Multidimensional scaling plot of shape examples using the inner product of histogram vectors to construct a distance measure

4   Object Classification: Experimental Results

We have tested our method on a database containing four categories of images: airplanes, cars, faces and motorcycles. A set of background images, that depicts varying objects, is also used. This database is the same as the one used in [6] downloaded from the image database of the Visual Geometry Group of the Robotics Research Group at Oxford at: `http://www.robots.ox.ac.uk/~vgg/data/`

Examples of images from the different sets are shown in figure 7, together with gradient images. The gradient images were obtained by gradient detection and subsampling after smoothing. By varying thresholds and subsampling rates we fixed the number of gradients to 500. Note that this occasionally implies that weak gradients from smooth areas in the background are detected. The fact that gradient strength is used in weighting the final index histograms serves to reduce the effect of these weak "noise" gradients.

Following the procedure in [6], we apply our method on each category in the following way. First the object set is split randomly into two sets of equal size. One set is for training; the other is for testing. The background set is used solely for testing. Thus there is no training of background images at this stage. For each image in the two test sets (object and background), the proximity to all the images in the training set is computed by inner multiplication of their index
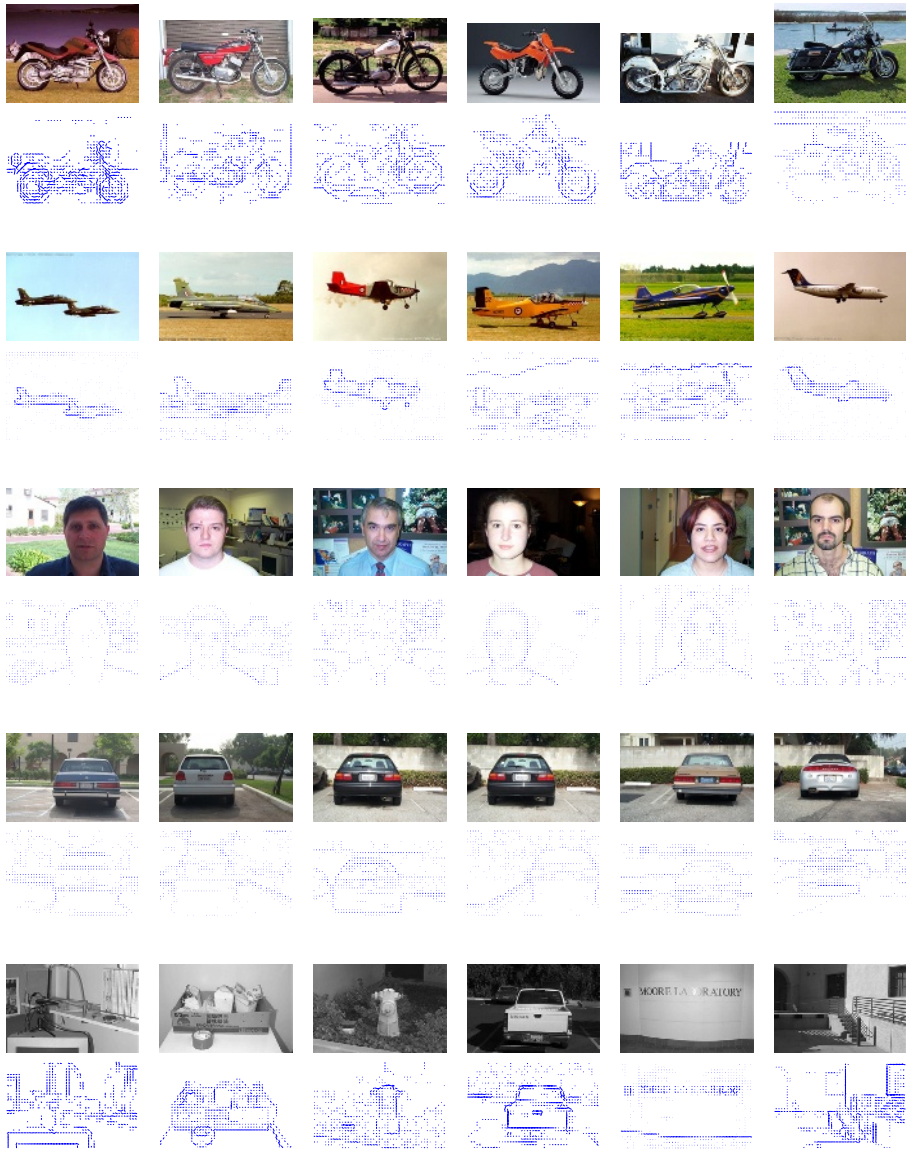
Fig. 7. Examples of images and gradient plots from the categories: airplanes, cars, faces, motorcycles, and background. These sets can be downloaded from the image database of the Visual Geometry Group of the Robotics Research Group at Oxford at: `http://www.robots.ox.ac.uk/~vgg/data/`

Table 1. Equal error rates for whole images.

Dataset	#Images	NN	Edited NN	Ref:[6]	Ref:[12]	Ref:[13]
Faces	450	81.8	83.1	96.4	-	94
Airplanes	1074	82.9	83.8	90.2	68	-
Motorcycles	826	93.0	93.2	92.6	84	-
Cars	126	87.3	90.2	84.8	-	-

histograms. The highest value, corresponding to the nearest neighbor (NN) is found for each element in the test sets.

To decide whether an image belongs to a category or not, we compare its nearest neighbor proximity value to a threshold. The receiver-operating characteristic (ROC) curve is formed by varying this threshold. The curve shows false positives (fp) as a function of true positives (tp). In figure 8 we can see the ROC curve for faces, airplanes, motorcycles and cars. The equal error rate (EER) is the rate where tp=1-fp. In table 1 we can see the equal error rates for the four different categories compared to results for the same categories achieved by [6]. In the 'Edited' column, we can see the values for the edited nearest neighbor method, where training is performed with background images as well. Using this method the background database is split randomly into two halves; one training set and one test set. The proximities between all histograms of both the object and background training sets are computed. For each histogram of the object training set, the k (typically 3) nearest neighbors are found. If all of those are from the object training set, the object histogram is kept; otherwise it is removed from the training set. This achieves smoother decision boundaries.

In table 1 we see that for two of the categories (cars and motorcycles) we get better results than [6], whereas for the other two categories (faces and airplanes) we get worse results. The worse results are mainly due to gradients from the parts of the image surrounding the object. They cause a lot of noise in the histograms, which lowers the proximity value of the nearest neighbor for true positives. This is confirmed by the fact that when objects in the image are cropped, results improve. This is seen in table 2. An example of a whole image versus cropped can be seen in figure 9.

Table 2. Equal error rates for cropped images.

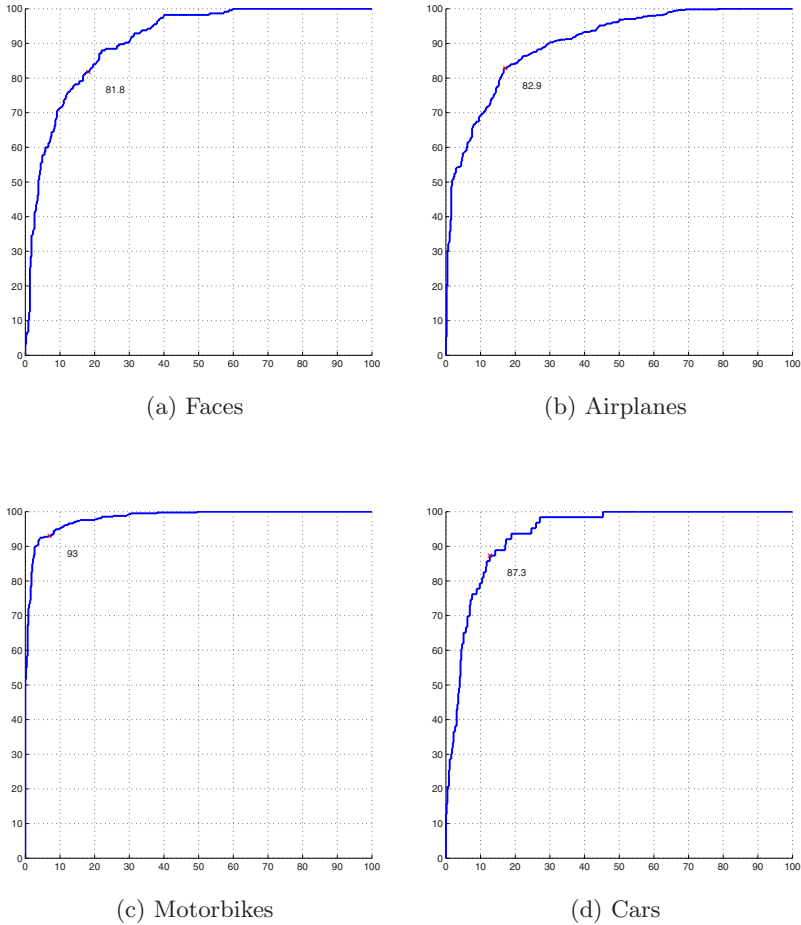| Dataset	#Images	NN	Edited NN
Faces	450	96.4	96.0
Airplanes	1074	86.7	89.2
Motorcycles	826	94.9	94.9
Cars	126	92.2	97.8

Fig. 8. ROC curves with EER for faces, airplanes, motorcycles and cars.

5   Summary and Conclusions

We have presented a simple method for qualitative image description based on histograms of order type + polarity computed from the locations and gradient directions of triplets of points in an image. We have demonstrated that this descriptor captures the similarity of images of objects within a class while discriminating against images of other objects and arbitrary backgrounds. We have avoided the resort to small image patches or fragments that is necessary for most image similarity methods due to their inability to capture large variations within a class. The experimental results on recognition of images four objects in a large database was in two cases better than the best results reported so far [6] which
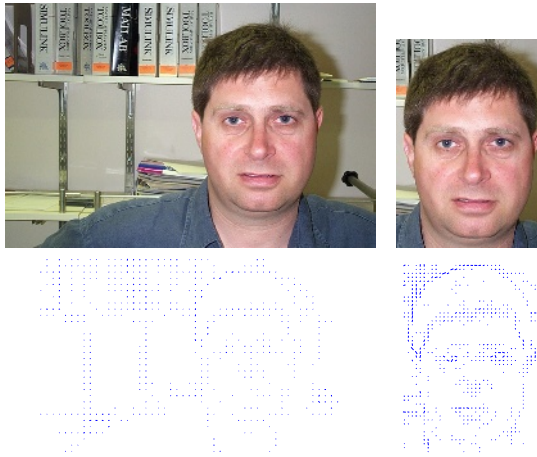
Fig. 9. Examples of whole and cropped face images and gradient plots.

uses a combination of patches to describe the object. In the two other cases, the amount of background clutter relative to the object was too large, giving results inferior to those in [6]. Using cropped images of these objects, our results are vastly improved demonstrating that by combining our algorithm with a search for the optimal size and location of a window covering the object, we should be able to handle these cases too.

The use of fragments can be motivated for reasons of recognition of occluded objects and for a potentially richer description of the object. In the future we will consider such descriptors where the size flexibility of our method will be of importance.

The use of qualitative image structure based on gradient locations and directions is essentially an appearance based method. It can however be tuned towards paying more attention to the strong structural properties of images which are undoubtly important in object classification, thereby forming a bridge between appearance and structure based methods for image recognition.

References

1. Belongie S. and Malik J. Matching with Shape Contexts Proc. 8:th International Conference on Computer Vision (ICCV 2001)
2. Borenstein, E., Ullman, S., Class-Specific, Top-Down Segmentation, Proc. ECCV 2002
3. Burl, Weber, Perona A probabilistic approach to object recognition Proc. ECCV 1998 pp. 628 - 641
4. Carlsson. S, "Order Structure, Correspondence and Shape Based Categories", *International Workshop on Shape Contour and Grouping,* Torre Artale, Sicily, May 26-30 1998, Springer LNCS 1681 (1999)

5. Carlsson S. and Sullivan J. Action Recognition by Shape Matching to Key Frames Workshop on Models versus Exemplars in Computer Vision, Kauai, Hawaii, USA December 14th, 2001
6. Fergus, R., Perona, P., Zisserman, A., Object class recognition by unsupervised scale-invariant learning, CVPR03(II: 264-271). IEEE Abstract. IEEE Top Reference. 0307 BibRef
7. Schiele B. and Crowley J. Recognition without Correspondence using Multidimensional Receptive Field Histograms, IJCV(36), No. 1, January 2000, pp. 31-50.
8. Schneiderman, H., A Statistical Approach to 3D Object Detection Applied to Faces and Cars, Proc. CVPR 2000.
9. Sung, K.K., and Poggio, T., Example-Based Learning for View-Based Human Face Detection, PAMI(20), No. 1, January 1998, pp. 39-51.
10. Thompson, D'Arcy, "On growth and form" Dover 1917
11. Thureson J. and Carlsson S. Finding Object Categories in Cluttered Images Using Minimal Shape Prototypes, Proc. Scandinavian Conference on Image Analysis, (1122-1129). 2003
12. Weber, M. Unsupervised Learning of Models for Visual Object Class Recognition, PhD thesis, California Institute of Technology, Pasadena,CA, 2000
13. Weber, M., Welling, M., Perona, P., Unsupervised Learning of Models for Visual Object Class Recognition, Proc. ECCV 2000, pp 18 - 32 Springer LNCS 1842