Appears in IEEE Conf. on Computer Vision and Pattern Recognition, San Diego, 2005.

Integrated learning of saliency, complex features, and object detectors from cluttered scenes

Dashan Gao Nuno Vasconcelos
Department of Electrical and Computer Engineering,
University of California, San Diego
dgao@ucsd.edu nuno@ece.ucsd.edu

Abstract

A novel procedure for object detection from cluttered scenes is proposed. It consists of an integrated solution to the problems of learning 1) a saliency detection module tuned to a class of objects of interest, 2) a set of complex features that achieves the optimal trade-off, in a minimum probability of error sense, between discrimination and generalization ability, and 3) a large-margin object detector. All stages of the new procedure have some degree of biological motivation and this is shown to enable a computationally efficient solution that is scalable to problems containing large numbers of object classes. Experimental evidence is given in support of the arguments that different levels of feature complexity are optimal for different object classes, and that optimal features range from parts to templates, depending on the variability of the object class.

1. Introduction

With the formulation of object detection and recognition as statistical classification problems and the advent of powerful classification architectures, the last decades have witnessed major improvements in detection and recognition accuracies. Yet, there are still various aspects in which the current state of understanding of these problems is too limited to allow the design of systems with the robustness and flexibility required by most practical applications. One of the significant limitations of current recognizers is a requirement for carefully controlled training, usually performed with large training sets that are manually assembled and pre-processed. This results in extremely lengthy data collection procedures that make it difficult to rapidly deploy a classifier for a given class of objects, if a training set is not already available for that class.

Lately, however, the vision community has started to investigate a new formulation of the detection/recognition problem, usually referred to as recognition from cluttered scenes, where it is assumed that the training examples are not previously segmented [2, 5, 4, 9, 10]. For example, a training set of faces will contain images where the faces are shown in front of some background scene that occupies the bulk of the image area. One aspect that makes the new formulation fundamentally different from the traditional, uncluttered, learning problem is the very unbalanced nature of the available example labels. While in the "negative" class every image region can be confidently assumed to be a "negative" example, for the "positive" class the situation is quite different. In fact, while each training image in this class is labeled as containing the object of interest, it is not clear which image regions are really "positive" or "negative" examples. This implies that every image neighborhood could potentially be of interest for learning, and leads to a potentially very large (and noisy) training set. Hence, in addition to the standard problems in detection and recognition (how to find good features, how many should be used, how to design an effective classifier) recognition from cluttered scenes requires the ability to learn which regions of each training image are informative for the task at hand, namely which regions contain the objects of interest.

This can be seen as a saliency problem, i.e. the problem of determining the image regions that are salient for detection/recognition purposes. Given a reasonable saliency module, it should be possible to extract a set of image regions containing the objects of interest, and then apply to this training set (complemented with a set of negative examples which are usually easy to find) any of the existing procedures for the design of object detectors or recognizers. Overall, the problem has two major components: 1) the identification of training examples and 2) the design of the classifier itself. Given that neither the saliency nor the classification stage are likely to be perfect, it appears that significant gains might be possible by integrating the two stages. The classifier should certainly improve when saliency is more accurate (because it will have access to a cleaner training set) and the saliency stage should be able to improve with feedback from the classifier (regarding image regions that it considered salient but were clearly identified by the classifier as not containing the object of interest).

This is the problem addressed by this work, where we present an integrated solution for saliency and classification in the context of object detection problems. The work includes various contributions that address significant open questions for this problem. The first is a discriminant formulation of saliency that is optimal in a classification sense and produces saliency locations which are most informative in the sense of identifying the object of interest. The second is an iterative procedure that relies on classification results to improve saliency, and on the improved saliency results to obtain a better classifier. The third is a procedure to generate a hierarchy of features of increasing complexity, which allows fine control over the trade-off between discriminant power (which increases with complexity) and generalization ability (which tends to decrease with increasing complexity). The fourth is a biologically inspired, and computationally efficient, mechanism for feature selection from overcomplete feature sets, that balances discrimination and redundancy reduction.

Overall, our results show that it is possible to simultaneously learn, in a strictly discriminant fashion, 1) a saliency detection module tuned to the object class of interest, 2) a set of complex features that achieve an optimal trade-off between discrimination and generalization for the detection of objects in that class, and 3) a large-margin object detector. It is also shown that different levels of feature complexity are optimal for different object classes, and that the optimal features range from parts to templates, depending on the variability of the class. All stages of our algorithm have some degree of biological motivation and this is shown to enable a computationally efficient solution that is scalable to problems containing large numbers of object classes, without compromising optimality in a classification sense.

2. Related work

Learning to segment and recognize objects from cluttered scenes is a topic that has received an increased amount of attention in recent years [1, 2, 5, 4, 9, 10]. A common theme to current approaches to this problem is to represent an object as a collection of parts. This leads to two fundamental questions: how to extract these parts from cluttered images, and how to represent them. The first problem is usually solved in one of two ways. The first is to randomly crop image patches from the images in the training set, at a wide range of scales, and then select those which are informative with respect to the object class [4]. This is a strategy of least commitment which guarantees that no fundamentally important patches will be lost due to coarse sampling. On the other hand, in order to guarantee coverage of the object of interest in its entirety, this approach usually requires sampling a very large number of image locations. This makes the subsequent step of patch selection computationally intensive and, so far, this method has only been demonstrated on databases of small images.

An alternative approach is to rely on a saliency detector to find a set of "interest points" in each training image [2, 5, 9, 10]. While drastically more efficient, from a computational point of view, this approach has weaker performance guarantees from an accuracy standpoint, because the definitions of saliency in current used are unrelated to the detection problem. Instead, saliency is usually defined as some universal property that salient image regions must exhibit. Particularly popular definitions are that the image region must 1) contain specific visual attributes, such as edges or corners [6], or 2) exhibit a significant amount of complexity, where complexity can be defined in multiple ways [8, 7]. Since these definitions do not constrain salient regions to be informative with regards the detection problem (e.g. are not tuned in any form to the class of objects to be detected) they tend to produce a collection of interest points that are only weakly guaranteed to have any relation to the object of interest. This increases the difficulty of the design of representative object parts, which has to be very robust to the presence of training outliers. The complexity is, in this way, shifted to the representation stage, which tends to be computationally intensive for these methods.

With respect to the representation of object parts, while some have argued for the use of simple features (e.g. local descriptors such as PCA, or SIFT [7]) [5, 9, 10], others have proposed complex ones (image patches) [2, 4]. Being more closely tuned to the objects of interest, complex features are certainly more discriminant. On the other hand, the response of simple features tends to exhibit less variability when images are subject to spatial image deformations, noise, or other perturbations. Overall, feature complexity is, for object detection, the main variable for controlling the trade-off between discriminant power and generalization ability (invariance) faced by any classifier. It therefore appears that best results should stem from 1) considering a hierarchy of features that span the continuum from simple to complex, and 2) learning the appropriate level of feature complexity for each detection problem.

3 Integrated saliency and object detection

In this work, we address all problems discussed above by proposing an integrated solution for learning saliency maps, object detectors, and features. In particular we propose an iterative procedure, illustrated by Figure 1, consisting of the

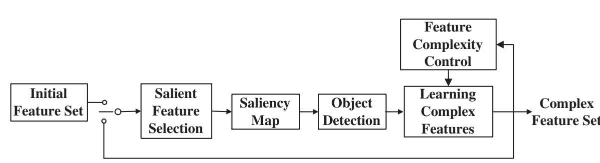


Figure 1. A hierarchical model for integrated learning of saliency maps, object detector, and complex features.

following steps. We start by selecting the most discriminant subset among a set of simple features (the discrete cosine transform - DCT - descriptors), which is used to generate a discriminant saliency map. Image patches are then extracted from the most salient locations and used to train an object detector. Using standard cross-validation the patches that are most likely to be positive examples are passed to a feature extraction module. The resulting features are more complex than the initial set and more tuned to the object class of interest. The process is then iterated and the complexity of the features allowed to increase at each iteration. The result is a feature hierarchy, that ranges from simple (DCT descriptors) to complex (image patches), allowing explicit control of the trade-off between discriminant power and generalization ability. A saliency map and an object detector are produced at each level of this hierarchy.

3.1. Salient feature selection

To avoid the lack of specificity of existing saliency detectors, we rely on a new formulation of saliency, denoted by *discriminant saliency*, which is intrinsically grounded on the recognition problem [11]: it equates saliency to the search for the visual attributes that best distinguish a visual concept from all other concepts that may be of interest. This leads to the formulation of saliency as discriminant feature selection for the one-vs-all classification problem that opposes the target image class ($Y = 1$) to all others ($Y = 0$).

3.1.1 Discriminant saliency detection

As shown in [11], discriminant saliency detection can be implemented through the combination of a scalable feature selection module and a biologically inspired saliency architecture, as follows.

1. images are projected into a K-dimensional feature space, and the marginal distribution of each feature response under each class $P_{X_k|Y}(x|i), i \in \{0,1\}, k \in \{0,\dots,K-1\}$, is estimated by a histogram (24 bins were used in the experiments in this paper). The features are then sorted by descending marginal diversity,

$$\mathbf{md}(X_k) =< KL[P_{X_k|Y}(x|i)||P_{X_k}(x) >_Y \quad (1)$$

   where $< f(i) >_Y = \sum_{i=1}^{M} P_Y(i) f(i)$, and $KL[p||q] = \int p(x) \log \frac{p(x)}{q(x)} dx$ is the Kullback-Leibler divergence between p and q [12].
2. features which are discriminant because they are informative about the background class ($Y = 0$) but not the class of interest ($Y = 1$), i.e. $H(X_k|Y = 1) < H(X_k|Y = 0)$, or that have too small energy to allow reliable inferences, $Var(X_k) < T_v$, are eliminated.
3. the features of largest marginal diversity are selected as salient for the class of interest. The number of features that are salient for each class is determined through a cross-validation procedure, as described in section 3.2.1.
4. a saliency map $S_D(\mathbf{x})$ is generated by projecting the image into the subspace spanned by the salient features, and combining the resulting projections $R_i(\mathbf{x})$ according to $S_D(\mathbf{x}) = \sum_{i=1}^{n} \omega_i R_i^2(\mathbf{x})$. The spatial support of the salient feature with the strongest response at each location is selected to be the scale associated with that location.
5. Salient locations are determined by non-maximum suppression. The location of largest saliency and its spatial scale are first found, and all the neighbors of the location within a circle of diameter equal to this scale are then suppressed (set to zero). The process is iterated until all locations are either selected or suppressed.

The method is made scale adaptive by including features of different resolution in the candidate feature set.

3.1.2 Class-specific feature sets

Unlike simple generic features, such as wavelet or DCT filters, class-specific feature sets learned from training images tend to be highly over-complete. This implies that they contain subsets of features which are highly redundant.

Feature selection in the presence of strong dependencies is, computationally, a much more demanding process than when such dependencies are not present. In particular, accounting for dependencies requires either 1) modeling joint densities, a process that has exponential complexity in the order of the dependency sets, or 2) penalizing the training samples that are well explained by the previously selected features, as is done by boosting [13]. Our experience with various existing feature selection methods, from both camps, is that that they would significantly compromise the computational efficiency of discriminant saliency.

Furthermore, we would like to embed feature selection in the computation of the saliency map itself, as some dependencies may have great impact on the latter while others may be irrelevant. To achieve this goal, we propose a biologically inspired feature selection procedure that combines aspects of the two feature selection strategies. As in the case of the simple features, we start by ordering the feature set according to the marginal diversity. We then pick features sequentially, in a manner that maximizes discrimination but penalizes redundancies. This penalty is implemented with the prime biological mechanism for redundancy reduction, non-maximum suppression, resulting in an example-re-weighting method for selecting features.

The use of non-maximum suppression to penalize dependencies is probably best understood by considering Figure 2, which presents an example in the context of face detection. A set of features, shown in (b), is initially available in result of the MMD-based selection step. These features are highly discriminant but also redundant. A reference image, shown in (a), is first randomly selected and individual saliency maps produced for that image (shown in (c))

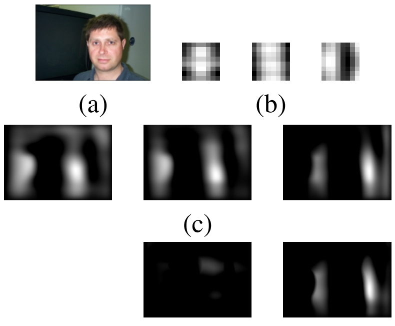


Figure 2. Illustration of non-maximum suppression. (a) A reference image. (b) Three features (c) Saliency maps (d) Saliency maps after suppression.

by considering one feature at a time. The largest response among these saliency maps is then found, and the corresponding feature selected. Non-maximum suppression then consists of subtracting the saliency map of the selected feature to all the others (shown in (d)). The process is iterated using the suppressed saliency maps until either 1) all features are selected, or 2) all the remaining saliency maps are below a threshold (set to zero in all results presented in this paper). As illustrated in Figure 2, the middle feature is highly redundant with the leftmost feature, and therefore discarded after non-maximum suppression. Overall, the proposed combination of feature selection and generation of the saliency map has great computational efficiency.

3.2 Learning complex features

The saliency maps of the previous section can be seen as soft segmentation masks for the object of interest. In this section we present a method that relies on these masks to learn complex features tuned to that object. Given that extracting complex features from areas not covered by the object of interest is likely not to be very useful, the extraction of complex features requires 1) determination of the best number of features to construct saliency maps, and 2) elimination of outlier locations in the resulting saliency maps. We refer to the combined process as the *extraction of representative object locations*. Once it is done, a collection of object patches can be obtained by retrieving the inlier salient locations, and a new set of features, more complex and tuned to the object, can be learned. We refer to this process as the *generation of complex features*.

3.2.1 Extraction of representative locations

We adopt a cross-validation strategy for the extraction of representative locations. The basic idea is to start from the saliency maps associated with images that contain the object of interest and extract image patches located at the points whose saliency is above a threshold (some examples are shown as circles in Figure 3). This produces a set of positive (object) examples. Repeating the process on images known not to contain the object produces a set of difficult negative (non-object) examples. The two sets are then used to learn a classifier. The process is repeated for all possible numbers of features used in saliency map design. By monitoring test error it is possible to determine the optimal value for this number.

The main difficulty associated with this procedure is that both the training and testing sets of the positive class are corrupted: because saliency is not perfect, they usually contain outlying background patches. To increase the robustness to this problem we adopt, as measure of classifier goodness, the probability of error on the task of classifying images, rather than patches. After all, this is the only data for which there is unmistakable ground truth. Images are classified in two steps. First, the image patches extracted from an image by the saliency detector are classified individually. Next, if at least one of the individual patches is classified as positive, the image is assigned to the object class. If all patches are negative, the image is assigned to the negative class. The ROC equal error rate (i.e. $p(Truepositive) = 1 - p(Falsepositive)$) of the resulting detector is used as a measure of the performance.

A support vector machine was used in the experiments, and the distance to the classification boundary was adopted as the measure for selecting the image patches, from the positive class, to be passed to the next stage. In particular only positive patches at a distance larger than the margin were selected. Some examples are shown in Figure 3.

3.2.2 Generation of complex features

The image patches extracted from representative locations are usually good prototypes for the object of interest. Nevertheless, because they are extracted from particular images, they tend to be too specifically tuned to the particular objects and viewing conditions captured by those images. It is, therefore, unlikely that they will generalize well if directly used as features for object detection [3]. Instead, good features must balance discrimination with robustness to variations in object appearance. As suggested in [3], one possibility to increase robustness is to reduce spatial resolution (or complexity). To generate features with a given level of complexity, we approximate the salient image patches by the best linear combination of a pre-specified number of simple features. In particular, if $\mathbf{I}$ is a salient image patch and $\{b_1, b_2, \ldots, b_N\}$ a set of N simple features, the best

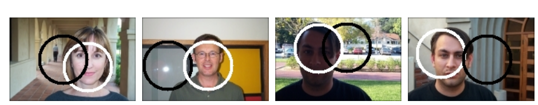


Figure 3. Examples of image patches accepted (white circles) and rejected (black circles) by the SVM.

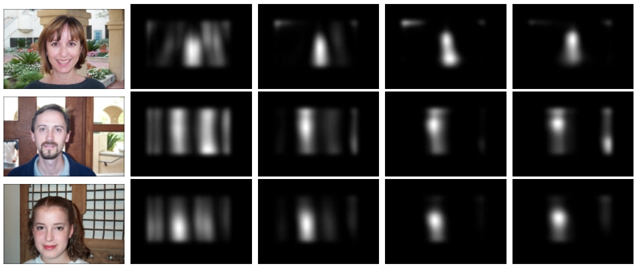
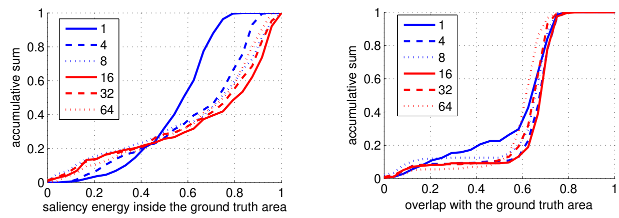


Figure 4. Saliency maps generated with features of different complexity ($k \in \{1, 4, 16, 64\}$ from the second to the last column).

Figure 5. Salient locations detected with features of different complexity ($k \in \{1, 4, 16, 64\}$ from the second to the last column). Circles are salient locations accepted (white) or rejected (black) by the SVM.



(a) (b)

Figure 6. Comparison between saliency maps, generated with features of different complexity, and the ground truth on the face database. Cumulative distribution of (a) percentage of salient energy inside the ground truth box, and (b) overlap between salient locations and ground truth.

approximation of complexity k is defined as the subset of k features whose linear combination best approximates the patch in the least squares sense

$$\min_{\substack{n_1 \dots n_k \\ a_{n_1} \dots a_{n_k}}} ||\mathbf{I} - \sum_{i=1}^{k} a_{n_i} b_{n_i}||^2, 1 \leq n_i \leq N \quad (2)$$

When the simple feature set is complete, the approximation error is monotonically decreasing on k and can always be made zero by making $k = N$. In this work we rely on a set of 8×8 DCT features to compose the simple feature set. This set is orthogonal and complete.

4 Results and discussion

The performance of the proposed object detection architecture was evaluated on the Caltech database, using the experimental set up proposed in [5].

4.1 Performance of complex salient features

Figure 4 shows some examples of saliency maps generated at different stages of the feature complexity hierarchy ($k \in \{1, 4, 16, 64\}$) for the face class. The simple and generic features used in the first stage appear to be sufficient for some scenes, but are not very selective for others, where they respond quite strongly to various areas of background. It is, however, clear that even for the simplest features the face regions always originate a strong response. At later stages, where the saliency is computed with complex features, the response is clearly stronger on the faces areas than the background, for all scenes.

To evaluate the saliency maps objectively we compared them, as well as the detected salient locations (in Figure 5), with ground truth, manually obtained by placing a rectangle around each face. The results are shown in Figure 6. The first measure is the percentage of the total energy of the saliency map that was contained in the ground truth box. Figure 6(a) presents the cumulative sum of this measure for features of different complexity. It can be observed that the saliency is more spread over the image for simple features than for complex features, confirming the observations of Figure 4. For example, the saliency maps generated by simple features have more than 70% of their energy inside the ground truth area 13% of the times, while for more complex features this percentage is always above 50%.

The second measure, whose cumulative sums are shown in Figure 6(b), is the relative overlap between the bounding box of the salient location and the ground truth. If A and B are two bounding boxes, the relative overlap is defined as

$$overlap(A, B) = \frac{|A \cap B|}{|A \cup B|} \quad (3)$$

where $|A|$ is the area of A. Again, complex features show better performance, but the differences are less significant.

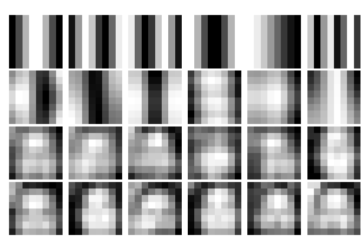


Figure 7. Six salient features generated at each stage of the feature complexity hierarchy ($k \in \{1, 4, 16, 64\}$ from top to bottom).

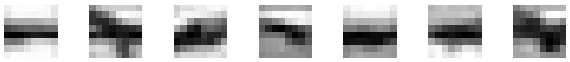


Figure 8. Seven of the salient features learned for airplanes by combinations of 32 DCT features.

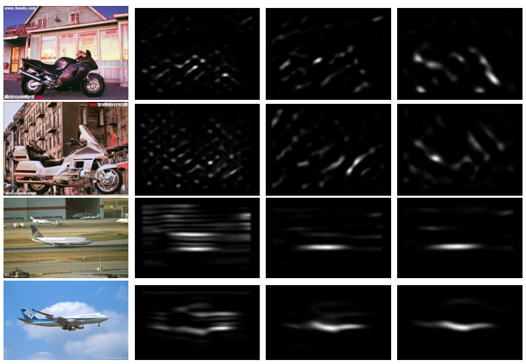


Figure 9. Saliency maps generated with features of different complexity, for the motorbike and airplane classes ($k \in \{1, 4, 16\}$ from the second to the last column).

This indicates that, even though the simple features respond more strongly in non-face areas, the strongest response is already quite reliably aligned with the face location. Overall, the best results are obtained with complex features composed by a linear combination of 16 simple features.

This result is probably best understood by investigating the salient features learned at each stage, as shown in Figure 7. In the first stage, salient features tend to be vertical bars, containing only very low frequency information about faces. As the complexity increases, and more high frequency information is added to the features, they start to look more like faces. In the final stages, where all 64 simple features are used, the features become face templates cropped from individual images. As discussed above, these templates are too tuned to individual faces and cannot account well for the variation inside the face class. In result, they lead to worse saliency maps than the ones of intermediate complexity.

We finalize the discussion on saliency by presenting, in Figure 9, saliency maps generated by features of complexity $k \in \{1, 4, 16\}$ for the motorbike and airplane classes. In general terms, the conclusions derived for the face class hold for these classes as well. The only significant difference is that, while for faces the learned complex features tend to be templates, for these classes they are parts, as can be seen from Figure 8. This shows that optimal complex features can range between the two types, depending on the variability of the object class.

4.2 Object detection

In this section, we evaluate the performance of the SVM classifiers at each stage of the hierarchy. Figure 10 presents

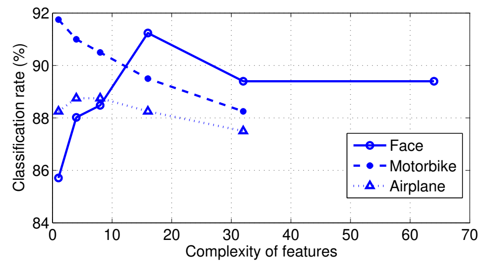


Figure 10. Classification rates for SVM designed at each stage.

the ROC equal error rates obtained at the different stages. For the face class, and consistently with the results of Figure 6, the best performance is achieved with the complex features of complexity $k = 16$. It is quite interesting, however, to realize that for the other two classes, motorbike and airplanes, simpler features actually work best. From the images of these two classes, shown in Figure 9, it is clear that these classes contain significantly more variability in appearance, pose, and scale than the faces. It is, therefore, not surprising that the performance of the complex features degrades in this case.

Another interesting observation is that, although there are mislabeled examples in the positive training set used to design the classifiers at all stages, these classifiers do not exhibit great difficulty in eliminating the mislabeled image patches and, consequently, generate good candidate features for the next stage. This is illustrated by Figure 3.

References

[1] Y. Amit and D. Geman, “A computational model for visual selection,” *Neural Computation*, Vol.11(7), pp.1691-1715, 1999

[2] S. Agarwal and D. Roth, “Learning a sparse representation for objection detection,” In *Proc. ECCV 2002*, Vol. 4, pp. 113-130, 2002.

[3] S. Ullman, M. Vidal-Naquet, and E. Sali, “Visual features of intermediate complexity and their use in classification,” *Nature Neuroscience*, Vol.5, No.7, pp. 1-6, 2002.

[4] E. Borenstein and S. Ullman “Learn to Segment,” In *Proc. ECCV*, pp. 315-328, 2004

[5] R. Fergus, P. Perona and A. Zisserman, “Object Class Recognition by Unsupervised Scale-Invariant Learning,” In *Proc. CVPR*, Vol. 2, pp. 264-271, 2003.

[6] C. Harris and M. Stephens, “A combined corner and edge detector,” *Alvey Vision Conference*, 1988.

[7] D.G. Lowe, “Object recognition from local scale-invariance features,” *Proc. ICCV*, pp. 1150-1157, 1999.

[8] T. Kadir and M.l Brady, “Scale, Saliency and Image Description,” *Int'l J. of Comp. Vis.*, Vol.45, No.2, pp. 83-105, 2001

[9] G. Dorko and C. Schmid “Selection of Scale-Invariant Parts for Object Class Recognition”, *Proc. ICCV*, pp.634-640, 2003

[10] A. Opelt, M. Fussenegger, A. Pinz and P. Auer, “Weak Hypotheses and Boosting for Generic Object Detection and Recognition”, *Proc. ECCV*, pp. 71-84, 2004

[11] D. Gao and N. Vasconcelos, “Discriminant Saliency for Visual Recognition from Cluttered Scenes,” *Proc. NIPS*, 2004

[12] N. Vasconcelos, “Feature Selection by Maximum Marginal Diversity,” *Proc. NIPS*, 2002

[13] P. Viola and M. Jones, “Robust real-time object detection,” *Second International Workshop on Statistical and Computational Theories of Vision Modeling, Learning, Computing and Sampling*, July 2001.