COMPLEX DISCRIMINANT FEATURES FOR OBJECT CLASSIFICATION

Sunhyoung Han   Nuno Vasconcelos

Department of Electrical and Computer Engineering
University of California, San Diego

ABSTRACT

A new algorithm for the design of complex features, to be used in the discriminant saliency approach to object classification, is presented. The algorithm consists of sequential rotations of an initial basis of simple features, so as to maximize the discriminant power of the feature set for image classification. Discrimination is measured in an information theoretic sense. The proposed algorithm has lower complexity than popular techniques for learning parts, and is evaluated on classification tasks from the PASCAL challenge. It is shown that complex features consistently outperform simple features.

Index Terms— feature selection, complex feature, visual recognition

1. INTRODUCTION

It has long been known that the careful selection of visual measurements, or *features*, is important for the solution of most image processing problems. In the area of object recognition, there has been a recent emphasis on localized representations, i.e. measurements that have a relatively small region of image support. This simplifies the design of the subsequent recognition stages, by constraining the dimensionality of the feature space in which they operate, and improves the robustness of the representation to geometric transformations, due to camera motion, pose variability, etc. There are two main types of localized representations, which we refer to as *features* [1] and *parts* [2, 3, 4, 5], and both have been widely used in the recent recognition literature.

Part-based representations rely on prototypical patches, usually produced by key-point detectors and clustering, which depict image-like structures. They have recently become quite popular for the representation of objects as constellations of parts [6] and image classification with "visual texture" [2, 4, 5]. These methods typically have significant computation, because the design of a part dictionary with good generalization requires learning a large codebook from a large number of training examples.

Since any orthonormal feature set spans the space of image neighborhoods of a given size, recognition can also be based on combinations of features that do not require learning, but simple selection of the best subset from a small number of orthonormal families (such as wavelets [7], Gabor [8, 9], or localized Fourier decompositions [1]). One low complexity solution of this type, first proposed in [10], relies on the principle of discriminant saliency. The idea is to, given a class of interest, find a set of features that are discriminant for that class (i.e. which best separate it from images in all other classes). Given an image to classify, salient locations can then be detected as the locations where the discriminant features produce a strong response. The resulting saliency map is indicative of the presence of objects from the class of interest in the image. Simple classifiers, whose input is this saliency, have been shown to achieve good classification performance, sometimes comparable to the state-of-the-art for much more complicated classification architectures, with minimal computation and significant robustness to clutter [10].

Previous work on discriminant saliency has relied on very basic methods for the selection of discriminant features. In this paper, we investigate the extension of the saliency framework, by providing it with the capability to learn the optimal feature sets. This is done through a computationally efficient feature extraction algorithm, which produces *complex* features. These are combinations of the original *simple* features, which are more tuned for the discrimination of the class of interest. The complex features now produced are more like the patches underlying the patch-based approaches, but can be learned with much less complexity. The performance of the new algorithm is tested on image classification problems from the PASCAL challenge [11], where it is shown that complex features can lead to improved performance.

2. FEATURE SELECTION

2.1. Simple features

The central problem for the design of a discriminant saliency detector is feature selection. Assuming that feature vectors are drawn from a random process $\mathbf{X} = (X_1, \ldots, X_n)^T$ according to a random variable $Y \in \{0, 1\}$ which determines the class ($Y = 1$ for objects in the class of interest, e.g. "faces", and $Y = 0$ for the null hypothesis, e.g. "non-faces"), the saliency of each feature is measured by the marginal mutual information between the feature and the class label [12]

$$I(X_k; Y) = < KL[P_{X_k|Y}(x|i)||P_{X_k}(x)] >_Y, \qquad (1)$$

where $KL[p||q] = \int p(x) \log \frac{p(x)}{q(x)} dx$ is the Kullback-Leibler divergence between the distributions $p(x)$ and $q(x)$ and $< f(i) >_Y = \sum_i P_Y(i) f(i)$. Salient features for the class of interest are those that maximize this mutual information.

2.2. Complex feature selection

The use of (1) makes feature selection tractable, from a computational point of view, but can limit the classification performance. Since the features are chosen independently of each other, any discriminant information which is captured by their dependences will be lost. In the feature selection literature, this problem is usually avoided by considering feature selection costs that account for such dependences. This, however, leads to an exponential increase in the complexity of the feature selection process. One alternative, which we pursue here, is to keep the cost of (1), but search for the feature space where this cost is sensible. This is done by selecting new basis functions Z_i, for the n-dimensional feature space, which are most discriminant than the initial X_i. Assuming that both the new and the existing basis are orthonormal, this can be achieved by searching for the rotation of the space which maximizes (1). The process is illustrated by figure 1, for $n = 2$. While, in the original space, it is impossible to achieve optimal classification by picking one of the X_i, the projection onto the rotated axis creates one optimally discriminant feature, and one which is completely non-informative for classification. The optimal rotation can be identified by searching for the space containing the feature which maximizes (1).

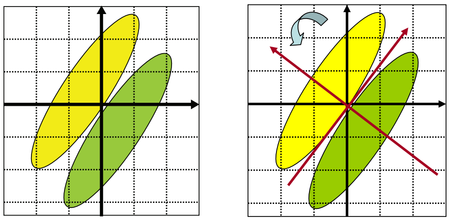

Fig. 1. Example two-dimensional classification problem, with two classes. Original basis (left) and rotated basis (right).

Let Z be the new feature, $Z = \phi^T \mathbf{X}$, where ϕ is a $1 \times n$ vector with $||\phi||^2 = 1$. The best projection, in a discriminant sense, is given by

$$\phi^* = \arg\max_{\phi} I(\phi^T \mathbf{X}; Y). \qquad (2)$$

However, the solution of this problem is still not trivial, since (2) has no closed-form solution. An exhaustive search for the optimal ϕ is also not feasible, given the high dimensionality of the space of rotations for any n of practical interest. To address this problem we adopt a coordinate-descent type of solution. Starting from the initial basis, we proceed iteratively, at each point identifying the subspace containing the two most discriminant features. We then find the best rotation within that subspace. Since this is a two-dimensional rotation, it can be performed efficiently by searching over a number of pre-defined rotation angles. If the old basis was orthonormal, the new basis is guaranteed to have this property, since the rotation takes place within a two dimensional subspace, which is orthogonal to all other dimensions of the space. The process is iterated until there is no increase in the sum of the marginal mutual information of (1).

The search, at each iteration, for the best two dimensional subspace can still be expensive. For example, if there are 64 features (8×8 image patches) there will be $\binom{64}{2} = 2016$ possible two dimensional subspaces. To improve efficiency, we restrict the search to those containing the currently most discriminating feature. This makes the search linear in the number of features, e.g. $O(63)$ in the example above. Since other features can always become most discriminant in subsequent iterations, we have found that this does not affect the feature selection results in any significant way. Overall, the algorithm is as follows:

1. set $\mathbf{\Phi} = \{\phi_i | i = 1, \ldots, n\}$, where ϕ_i is a $1 \times n$ vector of zeros except for a 1 in the i^{th} component ($\mathbf{\Phi}$ is the identity matrix)

2. find θ^* and j^* such that

$$\{\theta^*, j^*\} = \arg\max_{\theta, j} I(\cos\theta\phi_{i^*} X + \sin\theta\phi_j X; Y) \quad (3)$$

where i^* is the feature such that

$$\begin{aligned} i^* &= \arg\max_k I(\phi_k X; Y) \qquad (4) \\ &= < KL[P_{\phi_k X|Y}(x|y)||P_{\phi_k X}(x)] >_Y \end{aligned}$$

3. replace the i^{th} and j^{th} features with their rotation by θ^* i.e.

$$\begin{aligned} \phi'_i &= \cos\theta^*\phi_i + \sin\theta^*\phi_j \qquad (5) \\ \phi'_j &= -\sin\theta^*\phi_i + \cos\theta^*\phi_j. \end{aligned}$$

4. compute the overall mutual information

$$I = \sum_k I(\phi_k X; Y) \qquad (6)$$

go to 2) if it is larger than that of the previous iteration.

3. SALIENCY MAP GENERATION

Given an object class of interest, and an image where salient locations are to be identified, the saliency map is a map of weighted feature responses at all image locations. Each feature is weighted according to its discriminant power with respect to the classification problem that opposes the class of interest to the null hypothesis. The saliency $S(\mathbf{l})$ of location $\mathbf{l}$ is the weighted sum of the energy of all feature responses at that location

$$S(\mathbf{l}) \quad = \quad \sum_k I(X_k; Y) R_k(\mathbf{l})$$

where $R_k(\mathbf{l})$ is the result of half-wave rectification of the convolution of the image with the filter F_k, associated feature X_k [10]. We refer to $S(l)$ as the *saliency map* with respect to the class of interest.

4. SCALE ADJUSTMENT

Since the size of the object of interest, in the image where saliency must be determined, is usually not known, the saliency operation should search for the best image scale. This can be done by measuring feature responses at multiple scales, i.e. considering $\mathbf{X} = \{X_j^i | i = 1, \ldots, S, j = 1, \ldots, F\}$, where F is the selected number of features and S the number of scales, and searching for the scale

$$i^* = \arg\min_i \sum_{j=1}^{F} KL[P^t_{X_j^i}(x) || P_{X_j}(x)], \tag{7}$$

where $P_{X_j}(x)$ is the distribution of X_j in the training set (assumed to display images of roughly the same scale) and $P^t_{X_j^i}(x)$ the distribution on the test image of the responses of feature j and scale i. Feature responses of multiple scales can be obtained by applying the same feature set to various levels of a Gaussian pyramid decomposition of the test image.

5. EXPERIMENTS

To evaluate the impact of feature selection on discriminant saliency, we used an object classification task from the PASCAL challenge. A saliency map is produced for each image, histogrammed (in all experiments we used 36 bin histograms) and fed to a support vector machine (SVM). The SVM is trained to classify histograms into the class of interest and the null hypothesis. The saliency detector is evaluated by the accuracy of this classification.

5.1. Simple features vs. complex features

We start by analyzing the complex features produced by the proposed algorithm. We consider the Caltech face database, where the objects of interest (faces) have roughly constant size, and occupy a relatively large portion of each image. A comparison between the original simple features and the learned complex features is shown in figure 2. The number at the top of each feature indicates its scale (n means that the size of feature is 2^n by 2^n).

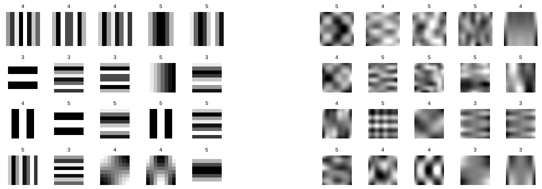

Fig. 2. Comparison of simple (left) and complex (right) features learned from the Caltech "face" class.

	simple	complex	scale simple	scale complex	[11]
bicycle	89.5	86.8	96.5	98.2	93
car	92	92	93.1	94.2	96.1
motorbike	92.6	93.5	94.9	94.9	97.7
people	91.7	97.6	95.5	97.3	91.7

Table 1. ROC equal error rate (detection rate at which the false positive rate is equal to the miss rate) for the four object classes on PASCAL.

Figure 3 presents a few examples of feature responses from the learned complex features. The first feature, whose responses appear in the first row, seems to capture the contour on the right side of the face. The second feature (second row) has strong response to the region around the left eye. The third feature (third row) appears to be tuned to the left half of the face.

5.2. Object category classification

For the classification experiments we relied on the PASCAL 2005 dataset 1. In this dataset, each image contains one from four classes of objects, plus background clutter. Table 1 shows the ROC equal error rate (EER) produced by the SVM histogram classifier, with various types of features (simple at single scale, complex at single scale, and the two types with scale selection). Overall, complex features achieve better rates than simple feature, and scale selection seems to be beneficial in both cases. The features learned for the "people" class are shown in figure 4. Note how the simple DCT features are transformed into complex 'face-like' features.

For completeness, we also present the best results reported in the literature (with more complex classifiers) for this dataset [11]. With complex features and scale selection, the simple classifier now proposed achieves better performance, than these methods, on two of the four classes. Figure 5 presents a comparison of the EER obtained with simple and complex features, as a function of the number of selected features. Note that, for all object classes, complex features produce better results, and the differences are larger when the number of features is small.

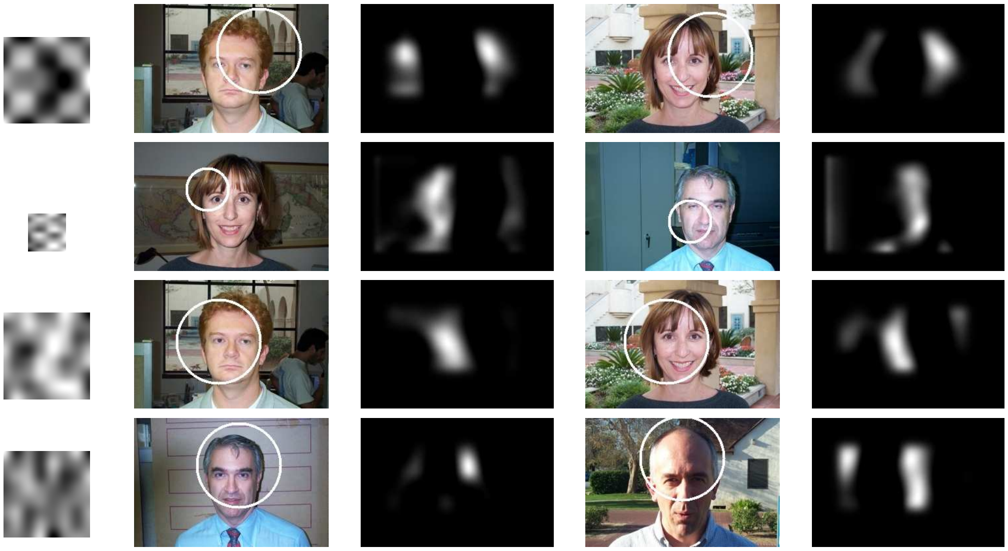

Fig. 3. Top four complex features and examples of their responses. Features are shown on the leftmost column, and responses to each feature fill the remainder of each row. In each case, we present the image on the left and saliency map (due to the feature only) on the right. The location of maximum response is highlighted with a circle of radius proportional to the scale of the feature.

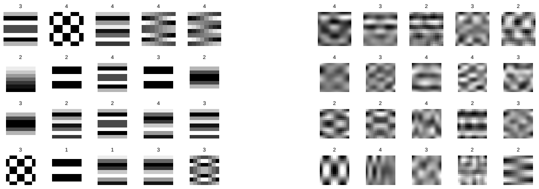

Fig. 4. Selected features. Simple DCT features(left) and complex features learned from the "people" class(right).

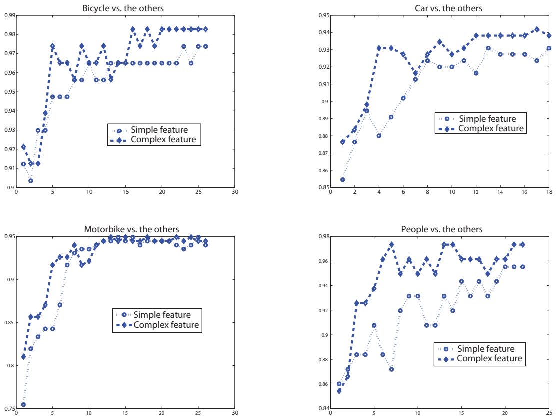

Fig. 5. ROC equal error rate according to number of feature. Objects are ”bicycle”, ”car”, ”motorbike”, and ”people” from the left top.

6. CONCLUSION

In this work, we have analyzed the impact of feature selection on discriminant saliency. Two conclusions can be drawn. First, complex features appear to improve object classification performance. Second, scale selection appears to be beneficial even with simple features. In the future, we plan to study how to account for variable scale during training.

7. REFERENCES

[1] N. Vasconcelos, "Minimum probability of error image retrieval," *IEEE Trans. Signal Processing*, vol. 52, no. 8, pp. 2322–2336, 2004.

[2] J. Mutch and D. Lowe, "Multiclass object recognition with sparse, localized features," in *Proc. IEEE Conf. CVPR*, 2006.

[3] G. Bouchard and B. Triggs, "Hierarchical part-based visual object categorization," in *Proc. IEEE Conf. CVPR*, 2006.

[4] G. Csurka and C. Dance and J. Willamowski and L. Fan and C. Bray, "Visual categorization with bags of keypoints," in *Proc. ECCV*, Prague, 2004.

[5] R. Fergus and P. Perona and A. Zisserman, "Object class recognition by unsupervised scale-invariant learning," in *Proc. IEEE Conf. CVPR*, 2003.

[6] S. Agarwal and A. Awan and Dan Roth, "Learning to detect objects in images via a sparse, part-based representation," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 26, no. 11, pp. 1475–1490, 2004.

[7] Paul Viola and Michael Jones, "Robust real-time face detection," *International Journal of Computer Vision*, vol. 57, no. 2, pp. 137–154, 2004.

[8] J.G. Daugman, "Complete discrete 2-d gabor transform by neural networks for image analysis and compression," *IEEE Trans. Acoustics, Speech, and Signal Processing*, vol. 36, no. 7, pp. 1169–1179, 1988.

[9] S. Mallat, "A theory for multiresolution signal decomposition: The wavelet representation," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 11, no. 7, pp. 674–693, 1989.

[10] D. Gao and N. Vasconcelos, "Discriminant saliency for visual recognition from cluttered scenes," in *Proc. NIPS*, pp., 481–488.

[11] The PASCAL Visual Object Classes Challenge 2005, "http://www.pascal-network.org/challenges/voc/voc2005/results.pdf," .

[12] N. Vasconcelos, "Feature selection by maximum marginal diversity," in *Proc. NIPS*, Vancouver, Canada, 2002.