Empirical Comparison between Hierarchical Fragments Based and Standard Model Based Object Recognition Systems

Sharat Chikkerur Lior Wolf

Abstract

The fragment-based architecture of Ullman's group [16, 12] which resulted in the recent work of Epshtein and Ullman [4] is a biologically motivated object recognition framework which bears similarities to the model of visual cortex developed in our lab [2, 14]. The fragment based method holds several advantages over our standard model approach such a top down part localization mechanism, minimal construction etc that are not currently available within the standard model. However, as far as we know, it has not been evaluated on standard datasets, nor compared to other existing systems. Also, there is no publicly available implementation of fragment based algorithms. Here, we provide such an implementation and compare the current instantiations of the two frameworks with regards to recognition capabilities on publicly available datasets. This work is not meant to capture the full potential of the two research directions. Rather, it is meant as a snapshot in time of their current capabilities with respect to their performance in object classification tasks.

1 Introduction

Physiological investigations have shown that object categorization in the primate visual systems is based on a hierarchy of features that progressively increase in complexity, ranging from simple line detectors to even entire objects (See [10, 13] for review). It is not surprising then, that biologically motivated object classification systems that use a similar hierarchical organization outperform holistic and part based counterparts [6, 1]. Some of the most recent hierarchical approaches to object recognition include Epshtein and Ullman's image fragment based technique [5] and Reisenhuber, Serre, Poggio et. al's [13] computational model of the visual cortex (the 'standard model'). In this memo, we present qualitative, as well as objective comparison between Epshtien and Ulmann's hierarchical fragment based classifier and the 'standard model'. While, Epshtein and Ulmann's system has shown promising results on some data sets, it has not been evaluated on publicly available standard data sets or objectively compared to other existing approaches. The standard model on the other hand was extensively compared to other front-end vision algorithms [14, 13, 11]. Here, we present a publicly available implementation of the fragment based classification system and perform empirical evaluation over standard datasets. We compare its performance with that of the standard model and a simple pixel based SVM classifier (used as an empirical measure of the data complexity).

2 Fragment Based Feature Hierarchies

2.1 Overview

In fragment based classifiers [9, 1, 16], the informative features in the images are represented by means of patches (fragments) taken from the training examples. The part based representation captures the commonly occuring object features, and at the same time accounts for intra-class variations. In [5], Epshtein and Ullman extend this approach by using a hierarchical architecture where each candidate fragment or feature derived from the training set is itself composed of informative sub-fragments. In the following sections, we outline

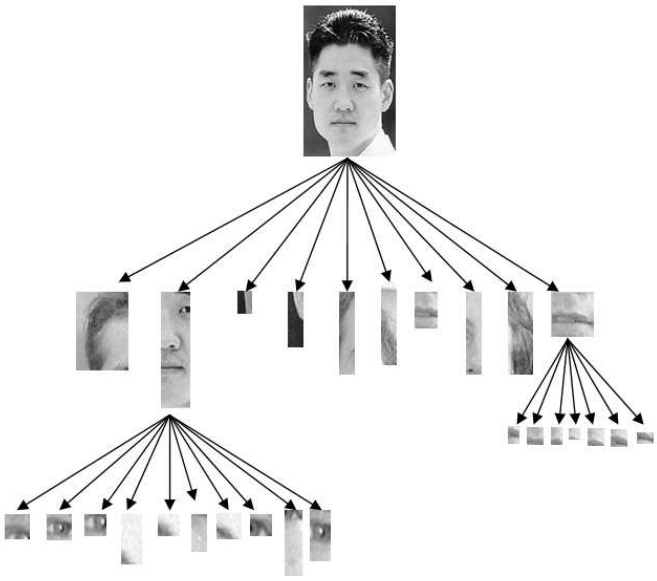

Figure 1: Fragment Hierarchy

their training and classification procedure in some detail. The reader is advised to refer to [4] for more details.

The algorithm takes as input positive and negative examples of the object categories and yields a feature hierarchy, such as the one shown in Figure 1. The training is divided into three distinct steps as outlined below.

1. **Feature Selection:** The features used for classification consists of image parts extracted from the set of training images. To begin with, the algorithm extracts several thousands of fragments of various sizes and from various positions within the positive and negative training examples. Among these, only a few fragments are selected such that the mutual information between the selected fragments and the class is maximized. In order to avoid evaluating mutual information for all possible subsets of fragments, a greedy approach is used. Once the most informative fragment is chosen, the next fragment is chosen to maximize the additional mutual information w.r.t the class while minimizing the mutual information with the already chosen fragments. The motivation and details about this mutual information driven feature selection process can be found in [8]. Once the optimal fragments are selected, a new set of positive and negative images is generated by selecting the regions where the fragments were detected in the original positive/negative training set. The most informative sub-fragments are generated by repeating the feature selection process using the new training data.

2. **Region of Interest(Receptive Field) Optimization:** In this step, each fragment and sub-fragment is associated with a receptive field or a region of interest, in which the fragment is assumed to be present. The region of interest has to be carefully chosen, since a very large receptive field leads to false positives and a very small receptive field will miss most feature occurences. Therefore, the receptive field is optimized to achieve a balance between these two extremes. The optimization is done by maximizing mutual information between the fragment and the class. Unlike the feature selection stage, the detection is based on results within the receptive field.

3. **Weight selection and classification:** The classification is done by overlaying a neural network over the fragment hierarchy (See Figure 2.1). Initially the weights are chosen to be a random value

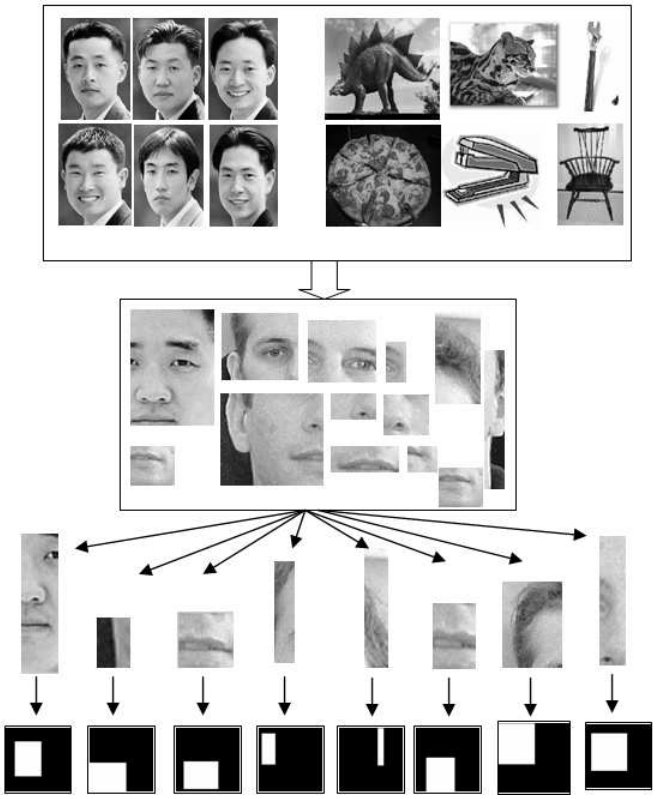

Figure 2: Overview of the training process

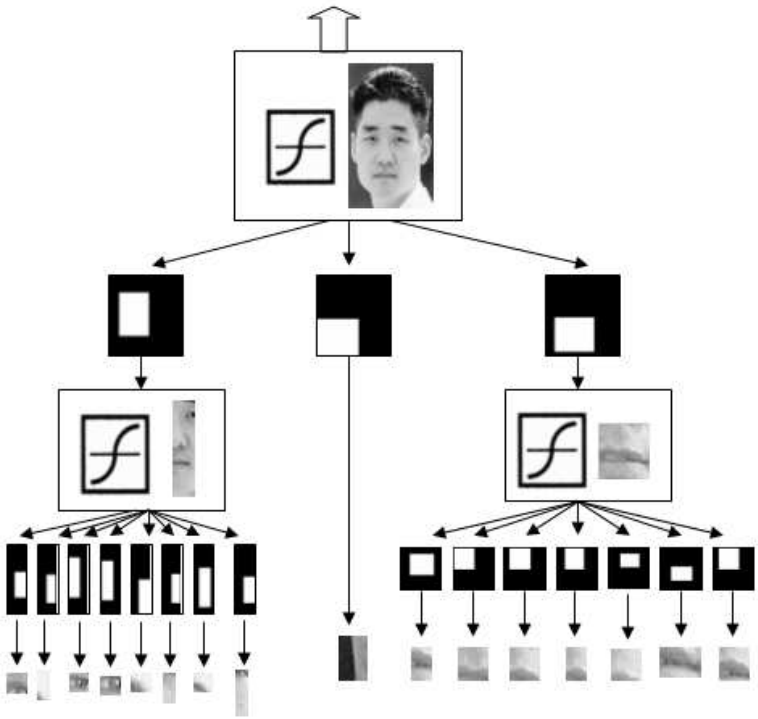

Figure 3: Overview of the classification process

between 0 and 1. Then, for each image in the training set, the response of the network is computed as follows. For each position of the parent fragment within its receptive field, a weighted response of its sub-fragments is evaluated. The response of the sub-fragment in turn, is computed as the maximum weighted response due to all its children within its receptive field (However, if the node is a leaf node, its response is considered to be the maximum normalized correlation measure within its receptive field). It is to be noted that the receptive fields are defined relative to the node's parents and therefore change with position of the parent fragment. The object is said to be detected if the final response at the root node crosses a certain threshold. Once all the reponses are computed, the weights are optimized using back-propagation algorithm. However, the response with these new weights are not the same as before, since the response at each level now depend upon the new weights. The weights and responses are iteratively optimized until the response converges. This usually occurs within five iterations in our implementation. It is also to be noted, that the weights are optimized together at all levels of the hierarchy.

3 Standard Model

3.1 Overview

The standard model provides a framework for computer vision that is consistent with biological evidence and in some cases even predicted it [13] . It is 'standard' because it attempts to consolidate all the widely accepted facts and observations about the primate visual system in a single model. However, it does not attempt to explain everything in the visual pathway. It has a narrow focus of explaining feed-forward mechanism in the first 100-150ms of the recognition phase. What happens after this or how the abundant feedback connections operate is not well known. But the following are widely accepted facts about the nature of object recognition in the primate visual cortex

1. The features and cells used for recognition are organized in the hierarchy with complexity and invariance increasing at each level (See [13] for review).
2. The processing sub units increase in number at each successive layer.
3. Rapid (100-150ms) object recognition does not involve any feedback mechanisms.
4. Learning occurs at each stage of the process.

The model is built up using a hierarchy of simple (S) and complex (C) units (here the term *unit* is used instead of *cell* to distinguish the entities in the standard model from their countepart in biology). The structure mimics similar organization found in the visual cortex, originally discovered by the breakthrough work of Hubel and Wiesel [10]). In the biological visual system, the S1 cells respond selectively to small lines at specific orientation. The response of these cells have been shown to be very well approximated by Gabor filters tuned at specific orientation and frequencies [3]. Complex cells on the other hand pool inputs from several afferent S cells. In a similar fashion, the S units in the standard model provide selectivity to specific stimulus while the C units provide invariance by aggregating information from the afferent S units. The receptive field of the units increases as we go higher up in the hierarchy. A simplified version of the architecture is shown in Figure 4. Here the S1 units consists of a bank of simple gabor filters at several scales and orientations. The specific shape of the filters were designed based on physiological evidence [3]. The C1 units aggregate information from several S1 units within its receptive field using a simple max-like operation. This operation done by the C units is the key to the invariance exhibited by the standard model, since the C cell responds to the afferent pattern occuring anywhere within its receptive field. As the receptive field increases in the higher layers, the S units respond to increasing complex patterns and the C layers provide increasing invariance in position and to some extent in scale.

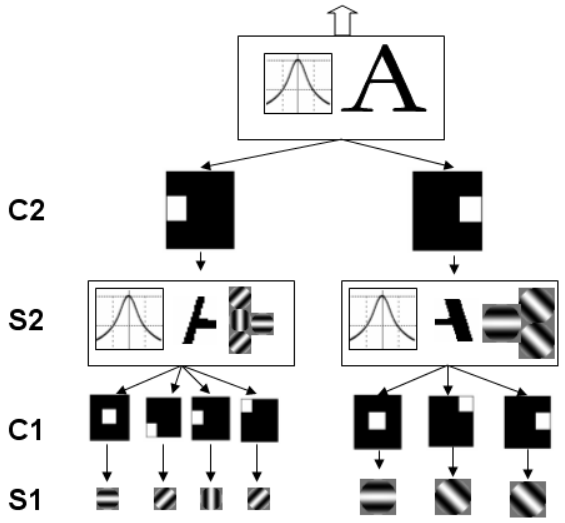

Figure 4: Standard Model Architecture

3.2 Training and Classification

For our comparison we use the simple two layer model used in [14] as our base-line for the standard model. During training S1 and C1 features are obtained by filtering the image with a bank of gabor filters and performing max operation on them. Instead of relying upon a universal dictionary of shapes [13] as in the more biologically precise model, the work in [15] relies upon learning object specific features in order to improve classification performance. After the C1 features are computed, random patches of various sizes are derived from the training images and form the S2 features that respond selectively to the object class. The S2 units have a gaussian like tuning function. The C2 cells again aggregate information from several S2 cells using a max operation. The C2 features (1000 in our current implementation) are used as a feature vector representating the class and used for training an SVM classifer.

4 Qualitative Comparison

4.1 Similarities

Figure 4.1 highlights the similarity between the two hierarchical models. Apart from the hierarchical organization of the two classifiers, there are other non-obvious similarities such as the following.

1. In most cases, the leaf nodes of hierarchy resemble simple edge, line and corner detectors. While the fragment based system explicitly learns these features from the training data, the standard model hardwires such simple detectors in the S1 layers with parameters that are consistent with physiological observations [14]. Therefore, if we compare the two models, the leaf nodes in the fragment based system represent the S1 layer of the standard model in a loose sense.

2. The fragment based model considers the response of a node as the maximum weighted response within its receptive field. As mentioned in [4], this directly emulates the HMAX behavior of the C layers in the standard model.

3. The parent fragments represent complex features composed of more simple features. The relative positions of the sub-fragments within their parent is captured by the receptive fields learnt during

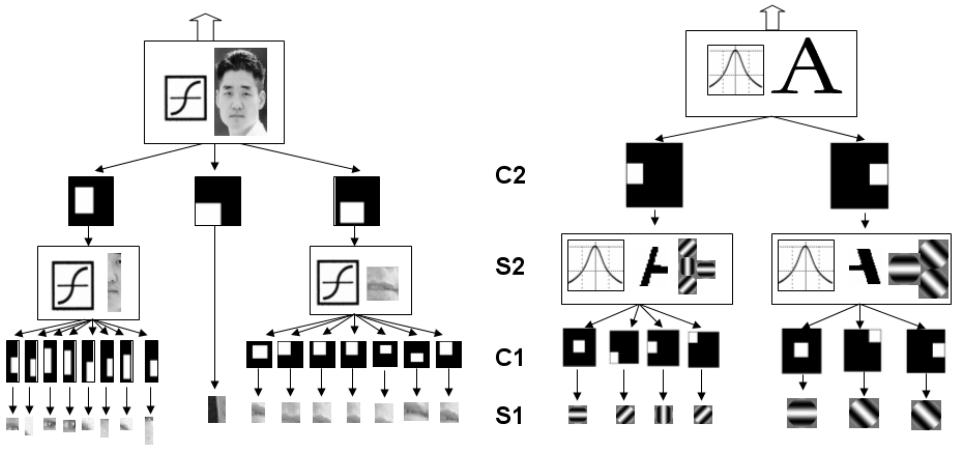

Figure 5: Qualitative Comparison

training. This can be compared with the complex S2 features that are derived by combining outputs of the afferent C1 layers, with the C1 layers directly modeling the receptive field.

4.2 Differences

The major differences between the two models are outlined in Table 1.

5 Experimental Comparison

In this section, we compare the performance of the two classifiers and also a simple pixel based SVM classifier that indicates to some extent the complexity of the task. We used images from the caltech database (http://www.vision.caltech.edu/html-files/archive.html) representing the standard data-sets and also from the Weizmann database (http://www.wisdom.weizmann.ac.il/ borise/hierarchies.htm) on which [4] results are based. The results reported are based on the best performance over three trials for each system. Summary statistics showing the minimum total error(False positive + False negatives) along with ROC area are given in Table 2

1. **Weizmann Faces:** This data-set consists of well segmented faces of caucasian and korean men. Out of these, we used 100 images for training and 100 for testing. The non-class images consisted of generic background images drawn from the caltech dataset.

2. **Weizmann Cows:** This data-set consists of segmented cow images. Out of these, we used 150 cows and non-cows images for training and the same number for testing. The non-class images consisted of an equal number of generic background images.

3. **Caltech Leaves:** The face dataset consists of 186 images of leaves from three different species taken against various backgrounds. Out of these, we used a random split of 89 training and 97 testing images. The non-class images consisted of 91 and 93 images an taken from a pool of generic background images. The splits in all the datasets that follow are consistent with those reported by [7].

4. **Caltech Faces:** This dataset consists of 450 images of about 27 individuals taken against various cluttered backgrounds. Out of these, we used a random split of 218 training and 218 testing images.

Fragment Hierarchy [4]	Standard Model [14]
Low level features are learnt from training examples	Low level features are hard-wired based on physiological data
Classification is based on a neural network model (at all levels)	The standard model based classifier used in this comparison [14], uses traditional SVM classifier (only at the last level)
Uses sigmoid like transfer function at the nodes, emulating a feed forward neural network	Uses an gaussian tuning function, emulating a radial basis network
Informative fragments are selected by maximizing mutual information	Informative fragments at the S2 layer are based on random selection of fragments of various sizes
Both the fragments and the receptive fields are learnt during training	Learning occurs only in the S2 layers. The receptive field size (size of the C1,C2 afferents) are fixed
The receptive fields are optimized by maximizing mutual information	The receptive fields are fixed
Feature locations can be detected using a top down approach	Currently, no such feature is present

Table 1: Major differences between the two models

The non-class images consisted of an equal number of images taken from a pool of generic background images.

5. **Caltech Motorcycles:** This dataset consists of 826 images of motorcycles taken from the side. Out of these, we used a random split of 401 training and 401 testing images. The non-class images consisted of 270 training and 228 testing images drawn from a pool of background images.

Database	Classifier	ROC area (range 0-1)	Min. error (range: 0-1)
Weizmann Faces	SVM	0.9900	0.00
	Fragment	0.9898	0.01
	Std. Model	0.9990	0.00
Weizmann Cows	SVM	0.9446	0.2067
	Fragment	0.9030	0.2000
	Std. Model	0.9650	0.1600
Caltech Leaves	SVM	0.9410	0.2541
	Fragment	0.8446	0.1156
	Std. Model	0.9820	0.0515
Caltech Faces	SVM	0.8593	0.3687
	Fragment	0.8919	0.2719
	Std. Model	0.9911	0.0533
Caltech Motor Cycles	SVM	0.9133	0.3228
	Fragment	0.8268	0.4412
	Std. Model	0.9993	0.0201

Table 2: Summary statistics for the empirical comparison

6 Conclusion

The qualitative compairson between the fragment hierarchy and the standard model shows that both the models share several traits in feature representation and classification. However, the fragment based classifier has several advantages due to its construction (at the cost of not being biological plausible) such as (i) minimal representation (as opposed to fixed hierarchy in the standard model), (ii) learning in both the feature detection and size of the receptive fields (as opposed to fixed receptive fields in the standard model), and (iii) ability to locate detected features using top-down tracing (the standard model, yields very little information apart from the object category after classification). With regards to performance, it can also be observed that it performs has high accuracy on well segmented datasets, but its performance degrades under more cluttered conditions (such as those in the caltech database). While minimal in construction, the training process is iterative and computationally expensive. The standard model on the other hand has an overcomplete representation, but also exhibits a higher level of invariance and robustness in cluttered conditions and consequently performs better than pixel based SVM classifier and part-based hierarchical classifier. The learning is fairly straight forward and computationally efficient compared to the hierarchical fragment based classifier.

7 Implementation

Both the standard model code and our implementation of the fragment based hierarchical classifier along with the instruction for use is freely available at http://cbcl.mit.edu/cbcl/software-datasets. The standard model may take about an hour to a couple of hours to train on a dual processor. Our current implementation of the fragment-based hierarchical model takes about a day or more to train on a dual processor (primarily to compute all the correlations). We have also posted matlab workspaces that contain the already trained hierarchies, weights and ROIs along with the positive and negative images and also the contexts they were

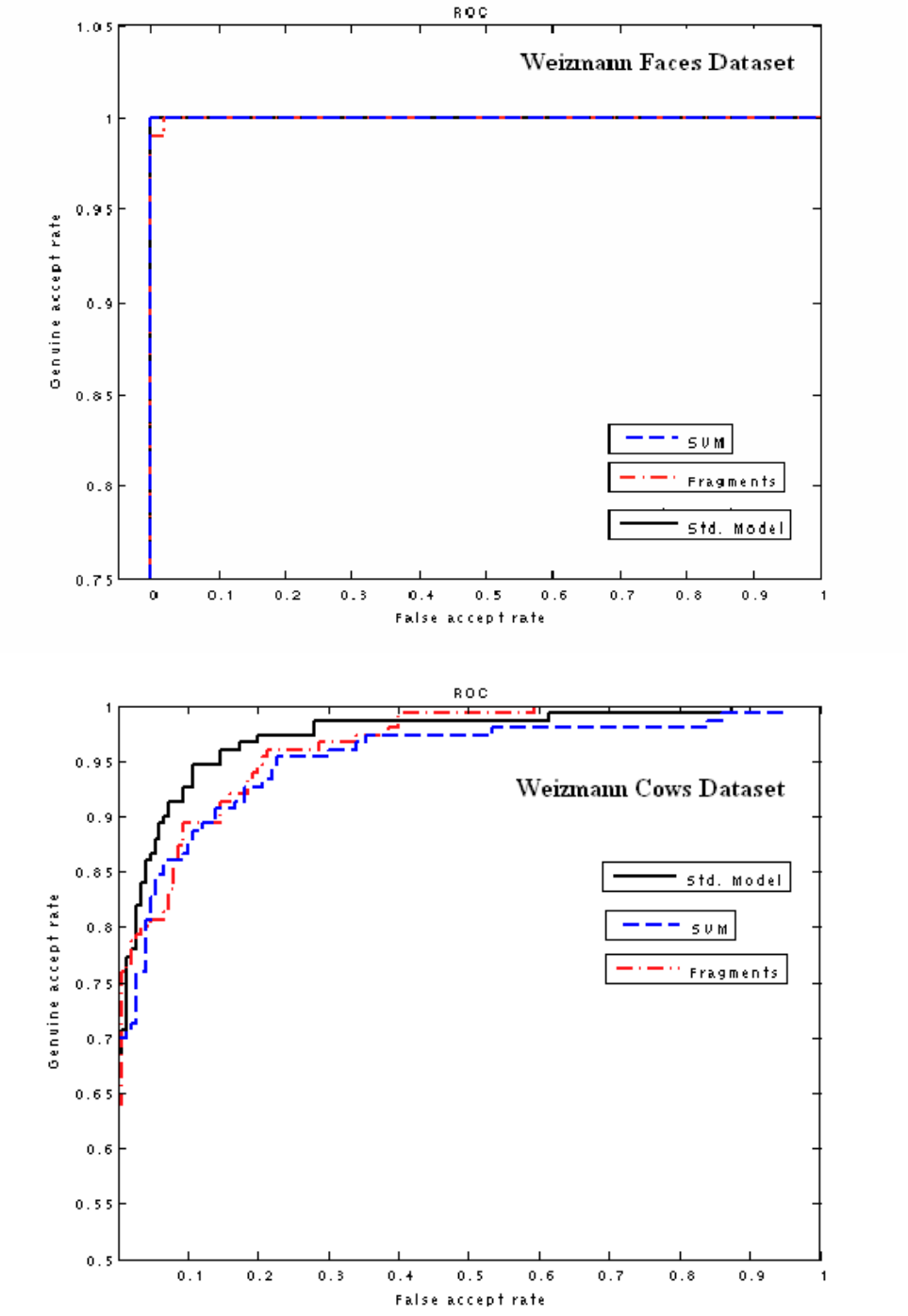

Figure 6: Comparative performance over Weizmann datasets

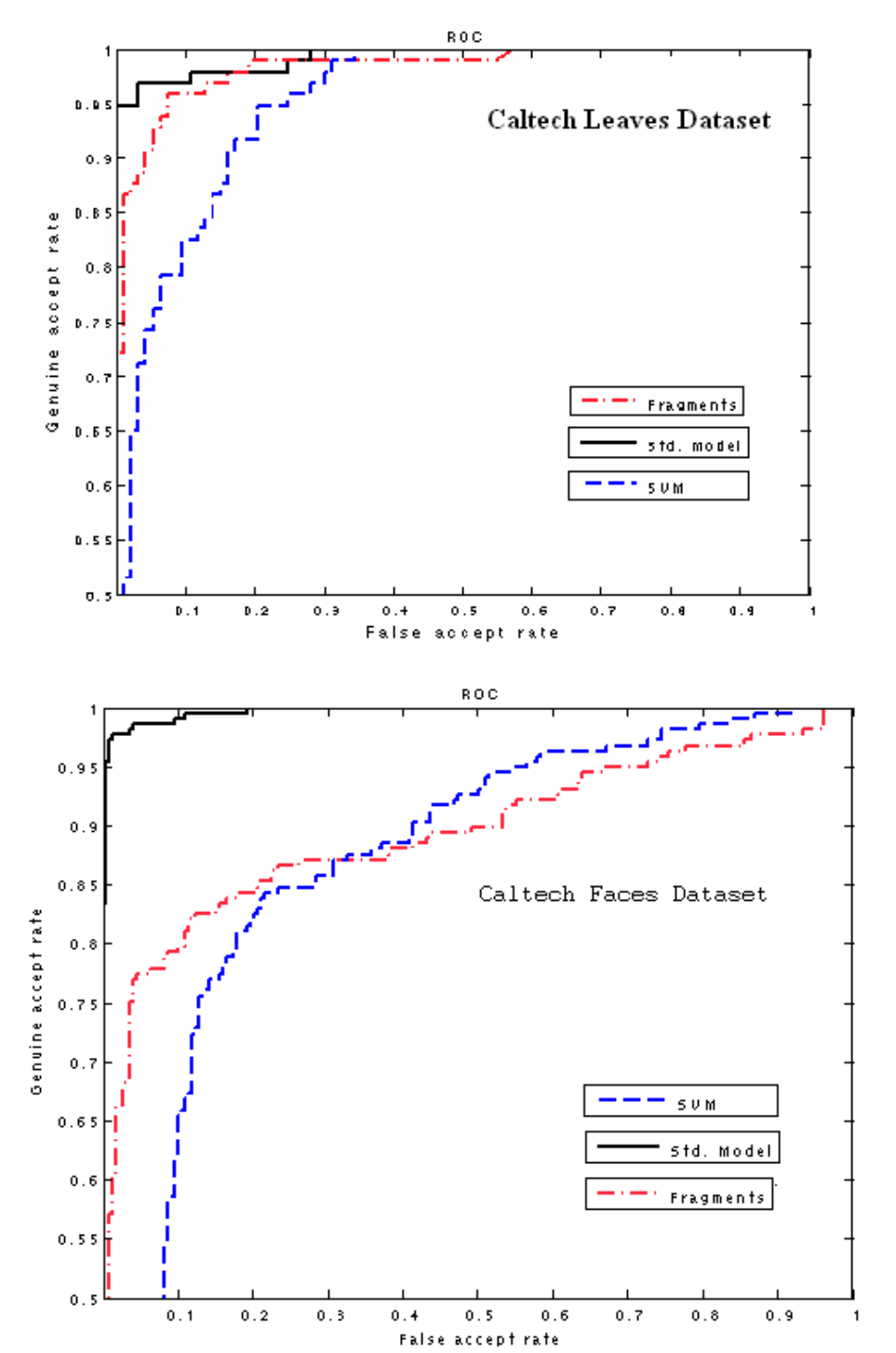

Figure 7: Comparative performance over Caltech datasets

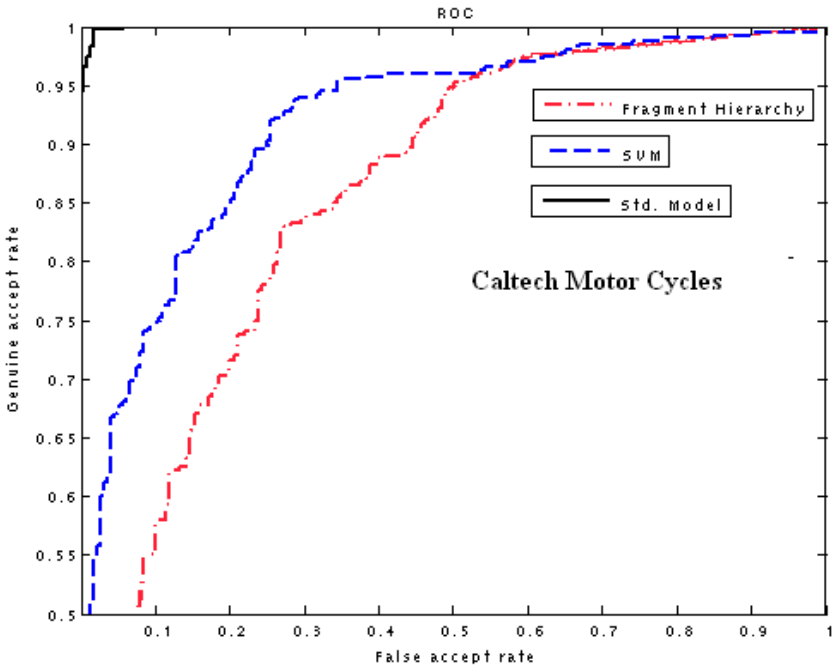

Figure 8: Comparative performance over Caltech datasets (contd.)

detected in at all levels of the hierarchy.

References

[1] S. Agarwal, A. Awan, and D. Roth. Learning to detect objects in images via a sparse part-based representation. *IEEE TPAMI*, 26(11):1475–1490, 2004.

[2] S. Bileschi and L. Wolf. A unified system for object detection, texture recognition and context analysis based on the standard model feature set. In *British Machine Vision Conference*, 2005.

[3] J. G. Daugman. Two-dimensional spectral analysis of cortical receptive field profile. *Vision Research*, 20:847–856, 1980.

[4] B. Epshtein and S. Ullman. Feature heirarchies for object classification. In *ICCV*, 2005.

[5] B. Epshtein and S. Ullman. Feature heirarchies for object classification. In *ICCV*, 2005.

[6] L. Fei-Fei, R. Fergus, and P. Perona. Learning generative visual models from few training examples: an incremental bayesian approach tested on 101 object categories. In *CVPR 2004, Workshop on Generative-Model Based Vision*, 2004.

[7] R. Fergus, P. Perona, and A. Zisserman. Object class recognition by unsupervised scale-invariant learning, 2003.

[8] F. Fleuret. Fast binary feature selection with conditional mutual information. *Journal of Machine Learning Research*, 5:1531–1555, 2004.

[9] B. Heisele, T. Serre, M. Pontil, T. Vetter, and T. Poggio. Categorization by learning and combining object parts. In *NIPS*, 2001.

[10] D. H. Hubel and T. N. Weisel. Receptive fields of single neurons in cat's striate cortex. *Journal of Physiology*, 148:574–591, 1959.

[11] J. Mutch and D. Lowe. Multiclass object recognition using sparse, localized hmax features. In *CVPR (under review)*, 2006. http://www.cs.ubc.ca/ mutch/.

[12] E. Sali and S. Ullman. Combining class-specific fragments for object classification. In *British Machine Vision Conference*, volume 1, pages 203–213, 1999.

[13] T. Serre, M. Kouh, C. Cadieu, U. Knoblich, G. Kreiman, and T. Poggio. A theory of object recognition: Computations and circuits in the feedforward path of the ventral stream in primate visual cortex. *MIT AI Memo*, 036, 2005.

[14] T. Serre, L. Wolf, and T. Poggio. A new biologically motivated framework for robust object recognition. In *CVPR*, 2005.

[15] T. Serre, L. Wolf, and T. Poggio. A new biologically motivated framework for robust object recognition. In *CVPR*, 2005.

[16] S. Ullman, E. Sali, and M. Vidal-Niquet. A fragment-based approach to object recognition and classification. In *IWVF4*, pages 85–100, 2001.