OBJECT RECOGNITION USING STRANGENESS AND TRANSDUCTION

by

Fayin Li
A Dissertation
Submitted to the
Graduate Faculty
of
George Mason University
in Partial Fulfillment of the
the Requirements for the Degree
of
Doctor of Philosophy
Computer Science

Committee:

__________________________  Jana Košecká, Dissertation Director
__________________________  Harry Wechsler, Dissertation Director
__________________________  Carlotta Domeniconi
__________________________  Zoran Duric
__________________________  Arun Sood, Chairman, Department
                               of Computer Science
__________________________  Lloyd J. Griffiths, Dean, School of
                              Information Technology and Engineering

Date: ______________________  Fall 2005
George Mason University
Fairfax, VA
Object Recognition Using Strangeness and Transduction

A dissertation submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy at George Mason University

By

Fayin Li
Bachelor of Science
Huazhong University of Science and Technology, 1996
Master of Science
Institute of Automation, Chinese Academy of Sciences, 1999

Directors: Jana Košecká and Harry Wechsler, Professors
Department of Computer Science

Fall 2005
George Mason University
Fairfax, VA
for my parents and brothers
Acknowledgments

I would like to thank my two advisors, Professor Jana Košecká and Professor Harry Wechsler, for all their help and advice during my studies at George Mason University. Their insights and enthusiasm have been a tremendous source of inspiration and have guided my path throughout these years. Their support, guidance, and patience have greatly benefitted me in research, writing, speaking, and many other aspects.

I am also grateful to Professor Zoran Duric, for his help, advising and research collaboration. I also wish to show my appreciation to Wei Zhang, Ho Shen-Shyang, all of the members of Computer Vision Lab. I also thank all my friends, near and far, for their support and care.
# Table of Contents

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Abstract</td>
<td>xi</td>
</tr>
<tr>
<td>1 Introduction</td>
<td>1</td>
</tr>
<tr>
<td>1.1 Motivation</td>
<td>1</td>
</tr>
<tr>
<td>1.2 Contributions</td>
<td>7</td>
</tr>
<tr>
<td>2 Transduction and Strangeness</td>
<td>8</td>
</tr>
<tr>
<td>2.1 Transductive inference</td>
<td>8</td>
</tr>
<tr>
<td>2.2 Strangeness measure</td>
<td>10</td>
</tr>
<tr>
<td>2.3 (k)-Nearest Neighbor Strangeness</td>
<td>15</td>
</tr>
<tr>
<td>3 Feature Detection, Representation and Recognition Approach</td>
<td>27</td>
</tr>
<tr>
<td>3.1 Feature Detection</td>
<td>28</td>
</tr>
<tr>
<td>3.2 Feature Description</td>
<td>31</td>
</tr>
<tr>
<td>3.3 Part-based Face Representation</td>
<td>33</td>
</tr>
<tr>
<td>3.4 Part-based Object Recognition Approach</td>
<td>36</td>
</tr>
<tr>
<td>4 Strangeness-based Feature Selection</td>
<td>41</td>
</tr>
<tr>
<td>4.1 Related Work</td>
<td>41</td>
</tr>
<tr>
<td>4.2 Variable Selection</td>
<td>43</td>
</tr>
<tr>
<td>4.3 Feature Instance Selection</td>
<td>49</td>
</tr>
<tr>
<td>4.4 Experiments and Evaluation</td>
<td>53</td>
</tr>
<tr>
<td>4.4.1 Variable Selection</td>
<td>53</td>
</tr>
<tr>
<td>4.4.2 Feature Instance Selection</td>
<td>53</td>
</tr>
<tr>
<td>4.5 Summary</td>
<td>57</td>
</tr>
<tr>
<td>5 Part-based Object Recognition with Strangeness</td>
<td>60</td>
</tr>
<tr>
<td>5.1 Related Work</td>
<td>60</td>
</tr>
<tr>
<td>5.2 Boosting Learning</td>
<td>62</td>
</tr>
<tr>
<td>5.3 Non-parametric weak learner</td>
<td>65</td>
</tr>
<tr>
<td>5.4 Learning Model and Recognition</td>
<td>69</td>
</tr>
<tr>
<td>5.4.1 Weakly Supervised Learning</td>
<td>71</td>
</tr>
</tbody>
</table>
5.5 Experimental Setup and Results ........................................... 73
  5.5.1 Location Recognition ................................................. 73
  5.5.2 Object Category Recognition ....................................... 75
  5.5.3 Face Recognition .................................................... 78
5.6 Discussion ........................................................................ 84

6 Open Set Recognition using Strangeness ............................... 86
  6.1 Background ..................................................................... 86
  6.2 Open Set Face Recognition ............................................. 87
    6.2.1 Open Set TCM-kNN Algorithm .................................. 90
  6.3 Experimental Results .................................................... 92
    6.3.1 Open Set Recognition Results .................................. 93
    6.3.2 Watch List ............................................................. 98
  6.4 Discussion ..................................................................... 100

7 Conclusion and Future Work ............................................... 103
  7.1 Conclusion ..................................................................... 103
  7.2 Future Work ................................................................... 105

Bibliography ............................................................................ 106
List of Tables

<table>
<thead>
<tr>
<th>Table</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>5.1</td>
<td>Recognition performance in term of percentage of correct localization.</td>
<td>73</td>
</tr>
<tr>
<td>5.2</td>
<td>The ROC equal error rates on the database used by Fergus et al. [18]</td>
<td>76</td>
</tr>
<tr>
<td>5.3</td>
<td>The performance of the final strong classifier at the second stage on the</td>
<td>78</td>
</tr>
<tr>
<td></td>
<td>database used in [18]</td>
<td></td>
</tr>
<tr>
<td>5.4</td>
<td>The performance of cars(rear), cougar faces and watches.</td>
<td>78</td>
</tr>
<tr>
<td>5.5</td>
<td>Top-1 rank performance with different learning approaches.</td>
<td>82</td>
</tr>
<tr>
<td>6.1</td>
<td>Mean Performance of Open Set PCA, Fisherfaces</td>
<td>99</td>
</tr>
<tr>
<td>6.2</td>
<td>Mean Performance of Open Set TCM-kNN.</td>
<td>100</td>
</tr>
</tbody>
</table>
## List of Figures

<table>
<thead>
<tr>
<th>Figure</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.1</td>
<td>Features detected in the training image. We need learn the object model from those features.</td>
</tr>
<tr>
<td>1.2</td>
<td>Local feature distributions for two classes.</td>
</tr>
<tr>
<td>1.3</td>
<td>Open Set Recognition.</td>
</tr>
<tr>
<td>2.1</td>
<td>Transduction [84]</td>
</tr>
<tr>
<td>2.2</td>
<td>The decision boundaries with different $N$ and $k$.</td>
</tr>
<tr>
<td>2.3</td>
<td>The decision boundaries with different $k$ when $N = 100$.</td>
</tr>
<tr>
<td>2.4</td>
<td>The decision boundaries with different $N$.</td>
</tr>
<tr>
<td>2.5</td>
<td>The test errors and their standard deviations with different $N$.</td>
</tr>
<tr>
<td>2.6</td>
<td>The relationship between strangeness and margin with different $N$.</td>
</tr>
<tr>
<td>2.7</td>
<td>The relationship between hypothesis-margin and sample-margin with different $N$.</td>
</tr>
<tr>
<td>2.8</td>
<td>The relationship between average strangeness and the posterior probabilities and odds with different training size $N$.</td>
</tr>
<tr>
<td>2.9</td>
<td>The boundaries constructed by kNN and strangeness with different $N$.</td>
</tr>
<tr>
<td>2.10</td>
<td>The test errors and their standard deviations with different $N$.</td>
</tr>
<tr>
<td>2.11</td>
<td>The test errors with respect to the dimensionality given the fixed $N$ training examples.</td>
</tr>
<tr>
<td>3.1</td>
<td>Affine Covariant Regions detected.</td>
</tr>
<tr>
<td>3.2</td>
<td>Difference-of-Gaussian Regions detected.</td>
</tr>
<tr>
<td>3.3</td>
<td>SIFT Feature Descriptor.</td>
</tr>
<tr>
<td>3.4</td>
<td>Golden ratio face template [4].</td>
</tr>
<tr>
<td>3.5</td>
<td>1st and 2nd local patches with initial scales.</td>
</tr>
<tr>
<td>3.6</td>
<td>The motivation of our approach.</td>
</tr>
<tr>
<td>3.7</td>
<td>Samples images of categories database contain faces, airplanes, cars, motorbikes and the background.</td>
</tr>
</tbody>
</table>
4.1 The accuracy achieved when using the features chosen by both algorithms. .......................................................... 54
4.2 The features in training data set and the results after feature selection. .................................................. 55
4.3 Feature instance selection results. .............................................................. 56
4.4 a) The total of 480 features detected and b) 50 discriminative features selected. .................................................. 57
4.5 The most likely selected features for each location. .............................................................. 58
4.6 The most likely discarded features by strangeness. .............................................................. 58
4.7 The original features detected (left) and the feature selected by strangeness (right). .................................................. 59
5.1 The object parts grouped – the weak rule in Boosting. ...................... 66
5.2 Boosting Learning with strangeness measure as the weak classifier. .... 70
5.3 Posterior $P(c|Q)$ Estimation for Query image $Q$ .......................... 71
5.4 Hierarchical Boosting Classification. .............................................................. 72
5.5 Examples of representative views of 12 out of 18 locations. .............. 74
5.6 Classification results with for Test Sequence 1 and Sequence 2 considering the spatial relationships modelled by HMM. The black circles correspond to the location labels assigned to individual frames of the video sequence. .............................................................. 75
5.7 (a) The ROC curve for image classification on the faces, motorbikes, airplanes and cars(side) data set used by Fergus et al. [18]. (b) The equal error rates with respect to the number of clusters $P$. ............ 77
5.8 Examples of the best feature for face after boosting learning. .......... 79
5.9 Examples of the best feature motorbike after boosting learning. .......... 80
5.10 Examples of misclassified image where few features are detected on the objects. .................................................. 80
5.11 Face Images .............................................................. 81
5.12 The average coefficients of the 1st order parts in final Boosting classifier. 82
5.13 The sample occluded images. .............................................................. 83
5.14 Average top-1 rank recognition rates with respect to $r$. ................. 83
5.15 Recognition rates with respect to $r$ when eye, nose or mouth is occluded. 84
6.1 Generic Open Set and Watch List Tasks. .............................................................. 86
Object recognition is one of the most essential functionalities of human vision. People can recognize large number of objects and easily provide for the “none of the known” answer when the unknown objects are present. It is of fundamental importance for machines to be able to learn and recognize meaningful objects and provide rejection option when the probe has no mates in the set of known objects. This thesis presents novel recognition algorithms using strangeness for open set recognition and generic object recognition, respectively.

Firstly, we present a supervised uncertainty measure, \textit{k-Nearest Neighbor strangeness}, to capture the difference of the data point from others with respect to class labels. The thesis quantitatively shows that strangeness can be used as a classifier to build the decision boundary and measure the uncertainty of the prediction. It also has one-to-one mapping relationship to the margin and the posterior probability. Based on those properties, we treat strangeness as the evaluation function and select the discriminative features from training data set to reduce the complexity of learning stage. The selected features enable the training data points to have large margin, which increases the probability of final classifier with good generalization capability.
Assuming that the objects are the random constellation of local visual parts which are common across the images in the same class while discriminative from other classes, we then employ strangeness as a non-parametric weak leaner in the boosting learning and present a novel part-based object recognition framework. In the framework, the class evidence of each part is characterized by strangeness and integrated by AdaBoost algorithm. The learning approach is able to handle the partial occlusion, local deformation, illumination, viewpoints and clutter background. It also provides the importance of each part for classification, which maybe helpful to understand what it means to be visually similar. The approach proposed is applicable for both two-class and multiclass classification problems. Its capability is demonstrated by supervised local recognition, part-based face recognition and weakly supervised object category recognition. For the later task, it can distinguish the object from both background and other objects, which extends the recognition capability much of current recognition methods for weakly supervised learning. The presented approach constitutes a significant step towards generic recognition and local evidence integration.

Finally, we present the Open Set TCM-kNN algorithm for open set recognition using strangeness. The algorithm provides a priori availability of a reject option to answer “none of the known” without modelling the distribution of any object and using negative training examples. The capability of the approach is demonstrated in a face recognition system. It provides a productive solution for open set recognition problem with small number training examples and large number of classes.
Chapter 1: Introduction

1.1 Motivation

Recognition is one of the most useful functions of our visual system. We recognize materials, surface properties, objects and scenes at a glance no matter the view angles, lighting illumination and local deformation. We recognize both individuals as well as categories. Even when we are very young, we can recognize thousands of categories and individual objects [11] and keep learning more throughout our life. A young child learns many categories and objects per day [11] without touching them. Given a category or object, we can easily judge if we have seen it before. Even for one category we know well (such as human, faces), we can quickly detect whether one object from the same category is “stranger” or not and provide for the “none of the known” answer without any negative training example, called open set recognition. Open set recognition is one crucial step for human incremental learning and acquire new knowledge. Replicating these abilities in the machines that surround us would profoundly affect the practical aspects of our lives, mostly for the better. This is one of the most exciting and difficult puzzle that faces computational vision, machine learning scientists and engineers in this decade.

A large number of diverse ideas has been proposed on object / category recognition while few work has been done for open set visual object recognition during the past few years. There is broad consensus of the fact that object models need
to capture the great diversity of forms and appearances of the objects, discriminate
the objects from others with partial information and characterize the statistics of
object appearances for rejection option. This means object models may be very
complex and contain hundreds, sometimes thousands, of parameters, which requires
a many-fold larger number of training examples [18, 77, 87, 89]. For the open set
recognition, FRVT2002 [29] used large number of faces to compute the statistics and
learn the threshold for rejection. In speaker verification, large number of negative
examples were used for scoring normalization in cohort or the universal background
model [26, 73].

Unfortunately, it is often difficult and expensive to acquire large sets of training
examples, especially for the negative examples in open set problem. Another problem
of most object recognition algorithms is that they require each training examples be
segmented and aligned (typically by hand). However image segmentation is NP-
complete problem [30]. Large number of pre-segmented training images require huge
human labor. Alignment may be further problematic for some recognition task when
fiducial points are not readily identifiable. This is a large, practical obstacle on the
way to learning thousands of objects. This thesis presents a strangeness measure
to address the following two recognition problems from small number of training
examples and represents a first step towards our long term goal of developing a generic
objects/categories recognition system:

1. Learning object/category models directly from images without segmentation
   and alignment.
2. Learning the rejection threshold for open set recognition without negative training examples.

The first problem is called weakly supervised learning, which is referred to the case when available training images contain both objects and varying background without prior segmentation. It aspires to emulate the visual learning ability of humans, recognizing that robust and generalizable solutions to practical recognition problems are likely to follow from a deep understanding of what it means to be visually similar. Weakly supervised learning is essential to advance the state of the art of current object recognition methods and scale them to be able to deal with large number of objects and/or object categories. In such a scenario, the objects are often represented by some number of local features to deal with partial occlusion, local deformation, clutter background, and object location and orientation robustly (see Chapter 3 for local feature detection and representation). Nevertheless the presence of large number of cluttered background features or irrelevant feature, as illustrated in Fig. 1.1, greatly increases the complexity of this task and prohibits many recognition algorithms to be used, which either require the clean and pre-segmented training images or assume that the training data for each class are sampled from some one underlying distribution. Even the best classification approach will fail without any preprocessing.

Given multiple training images for each class, the above problem can be generalized as following: Given a bag of relevant and irrelevant local features for each class, learn the model of objects, as illustrated in Fig. 1.2. Regarding this problem, many algorithms have been proposed recently [15,17,18,53,82], considering the appearance information of local features and/or the geometric spatial relationships. Among them,
Figure 1.1: Features detected in the training image. We need learn the object model from those features.

Figure 1.2: Local feature distributions for two classes.

most recently existing method

- Use voting approach. In this scheme, the recognition is determined by the number of matched features from the query image to model views in training database. The questions how to determine the threshold for matching? and how to estimate the matching confidence are still open problem. It only works well for rigid individual object recognition.

- Capture the uncertainty of descriptor and spatial relationship in a Bayesian
framework [17, 18, 33]. It only works with a small number of components to build the model of objects due to the computational complexity of learning. It only can classify the objects from background and has no discrimination ability among object categories.

- Using cluster technique to cluster all features into clusters [15, 82]. It was very sensitive the cluster algorithm and percentage of the irrelevant features. Large number of irrelevant features may have large number of clusters and overwhelm the relevant features in the clustering algorithm.

In a sense, all these techniques have their drawback and can be only applied in certain scenarios. Is there any learning framework working for both individual objects and object categories recognition in the presence of large number of irrelevant features? We believe so. In this thesis, we hypothesize that each kind of object is composed of several components which are common across the objects in the same class and discriminant from other classes, as illustrated in Fig. 1.2. Through the strangeness-based feature selection, the irrelevant features are discarded and discriminative features are selected, which build the components of the objects and yield the local evidence provided by individual features for recognition. General boosting strategy is used to integrate the local evidence, characterized by strangeness, and fuse them into final strong classifier. The whole framework is applicable for both individual objects and object categories recognition, and insensitive to the irrelevant features and cluster algorithm. It has not complex parameter estimation and can be used for individual objects and category recognition.

For open set recognition problem, it operates under the assumption that not all the
test probes have mates in the training data set and the imposters present, as shown in Fig. 1.3. It requires the a priori availability of a reject option to answer “none of the known”. A young child can easily detect the “unknown” objects. It seems unlikely that this would require a large set of training images for each class as well as negative training examples. Because of its difficulty, few methods have been proposed in past few years. Most of the existing methods use the negative examples to learning the a priori rejection threshold [26, 29, 73]. In this thesis, we use strangeness to investigate this problem in a transductive inference framework without modelling the distribution of any object and negative training examples. The comparative advantages of our proposed method come from its non-parametric implementation and automatic threshold selection. No assumptions are made regarding the underlying probability density functions responsible for the observed data. Our experiments show that this is a productive approach for open set recognition problem with small number training examples and large number of classes.
1.2 Contributions

Based on the strangeness, we proposed a general object recognition approach and a mechanism for open set recognition. The main scientific contribution of this thesis are:

- A new feature selection mechanism, which has the properties of both filter and wrapper approach. It has relatively low computational complexity and works even the training sample size is small.

- Incorporation of the new weak learner within the boosting learning framework, which can be used for both supervised object recognition and weakly supervised object/category recognition, even with small number of training images.

- Open Set TCM-kNN (Transduction Confidence Machine - k Nearest Neighbors) algorithm, which is suitable for multi-class authentication operational scenarios that have to include a rejection option for classes never enrolled in the gallery.

This thesis begins by describing transductive inference, giving the definition of strangeness and exploiting its properties (Chapter 2). It then describes how features are detected and represented from images (Chapter 3) and presents feature selection algorithms using strangeness (Chapter 4). A part-based object learning approach using strangeness is then proposed and demonstrated with different recognition tasks (Chapter 5). Chapter 6 further shows how strangeness can be used to build confidence machine and applied in open set recognition problem.
In Chapter 1, we introduce the problems met in generic object/categories recognition and open set recognition. In order to discard the irrelevant feature, integrate the local evidence and learn the rejection threshold, we need measure the uncertainty of the local features or object instances related to the class labels. In this chapter, a new informative measurement *Strangeness* is exploited to measure the uncertainty of the individual data points with respect to the class labels. The strangeness measure borrows the idea from transductive inference and can be used to construct confidence machine [68, 75] with randomness test [49] and predict the class with confidence and reliability (see Chapter 6). We demonstrate how the strangeness can be used to construct the decision boundary and examine its generalization capability. It can further be used to select discriminative features (see Chapter 4), integrate local evidence without complex parametric model learning (see Chapter 5).

### 2.1 Transductive inference

Transductive inference is a type of local inference that moves from particular to particular [83](see Fig. 2). “In contrast to inductive inference where one uses given empirical data to find the approximation of a functional dependency (the inductive step [that moves from particular to general]) and then uses the obtained approximation to evaluate the values of a function at the points of interest (the deductive
step [that moves from general to particular]), one estimates [using transduction] the values of a function [only] at the points of interest in one step” [83]. The simplest mathematical method for transductive inference is the method of k nearest neighbor classifier. The Cover-Hart [14] theorem proves that asymptotically the one nearest neighbor algorithm is bounded above by twice the Bayes minimum probability of error. Cover and Hart also showed that the k-NN error approaches the Bayes error (with factor 1) if $k = \Theta(\log n)$.

Figure 2.1: Transduction [84]

Vapnik [83] approaches transductive inference as follows. Given training (labeled) exemplars one seeks among feasible labeling of the (unlabeled probe) test exemplars the one that makes the error observed during testing (labeling) consistent with the error recorded during training. It is also assumed that the training and test exemplars are independent and identically distributed (i.i.d) according to some unknown distribution function. Vapnik defines then the overall risk functional using two equivalent settings with the explicit goal of minimizing them [83]. Setting #1 seeks to minimize the deviation between the risk on the training and working samples. Setting #2 labels $W$, the working exemplars set, in a fashion consistent with $T$, the training set,
e.g., using minimization for some error functional over $T \cup W$. An example is TSVM (Transductive support vector machines), which has been shown to yield substantial improvements for text classification, a two-class classification task, over inductive methods, especially for small training sets and large test sets. The observed success was explained due to the fact that “the margin of separating hyperplane is a natural way to encode prior knowledge for (text) classifiers and to study the location of text exemplars (with respect to the hyperplane), which is not possible for an inductive learner” [35].

The goal for inductive learning is to generalize for any future test set, while the goal for transductive inference is to make predictions for a specific working set. The working exemplars provide additional information about the distribution of data and their explicit inclusion in the problem formulation yields better generalization on problems with insufficient labeled points [27]. Transductive inference becomes suitable for our problems since we lack enough training data for inductive learning. In the following section, we exploit how strangeness is derived to measure the uncertainty of the individual data points with respect to the class labels.

2.2 Strangeness measure

There have been numerous attempts to assign probabilities to machine learning classifiers in order to interpret their decision with confidence and reliability over all possible classes. In fact, we can trivially convert every machine learning classifier’s output to a probability distribution by assigning the predicted class the probability between 0 and 1 to all possible classes. However, such estimation may in general not be good or can overfit due to inherent applied algorithm’s biases. Bayesian approach is often to
use to estimate the posterior probability of the predicted class, which is argued to be always the best method and without “overfitting” problem. The important issues in Bayesian method are how to choose a good prior distribution and how to efficiently calculate the posterior. However, the optimality of the Bayesian method is based on the assumption that the data we observe are generated according to one of the distribution models in the chosen class of models. While this assumption is attractive for theory, it almost never holds in practice. In practice, one usually uses relatively simple models, either because there is not enough data to estimate the “true” model and the computational complexity of estimation is prohibitive, or because our prior knowledge of the system is only partial. Even when very complex models are used, it is rarely the case that one can assume that the data are generated by a model in the class.

Instead of assuming a family of models, Vovk et al [58,68,75] introduce an individual strangeness measure and construct the confidence machine using the algorithmic theory of randomness and transduction methodology. They made a weak assumption of the model space, that the data is generated i.i.d. with some distribution $D$ but make absolutely no assumption about $D$ other than it is a fixed distribution. For each “significance level” $\delta$, the confidence machine can either predict the single label or some prediction region with confidence and reliability based on the strangeness measure. The produced confidence values is valid in the following sense. Given some possible label space $Y$, if an algorithm predicts some set of labels $R \subseteq Y$ with confidence $t$ for a new example which is truly labelled by $y \in Y$, then we expect the
following to hold over randomization of the training set and the new example:

\[ P(y \notin R) \leq 1 - t \quad (2.1) \]

Moreover, we prefer the algorithms which give “near precise” confidence values, that is values such that Equation 2.1 approaches equality. Now we give a brief outline of the confidence machine framework.

Consider a sequence of examples \((z_1, \cdots, z_n) = ((x_1, y_1), \cdots, (x_n, y_n))\) drawn independently from the same distribution over \(Z = X \times Y\) where \(Y\) is some label space. A family of functions \(\{A_n : n \in \mathbb{N}\}\), where \(A_n : Z^n \to \mathbb{R}^n\) for all \(n\), is called an individual strangeness measure if, for any \(n\), any permutation \(\pi : \{1, \cdots, n\} \to \{1, \cdots, n\}\), any \((z_1, \cdots, z_n) \in Z^n\) and any \((\alpha_1, \cdots, \alpha_n) \in \mathbb{R}^n\) it satisfies the following criterion [58]:

\[(\alpha_1, \cdots, \alpha_n) = A_n(z_1, \cdots, z_n) \Rightarrow (\alpha_{\pi(1)}, \cdots, \alpha_{\pi(n)}) = A_n(z_{\pi(1)}, \cdots, z_{\pi(n)}) \quad (2.2)\]

The meaning of this criterion is that the strangeness measure should be produced for each individual element in the data set, regardless of the order in which the values are calculated. Furthermore, the strangeness function must also have the following property. For every \(n\) and the output \((\alpha_1, \cdots, \alpha_n)\) of \(A_n\), there exists a semi-computable function \(t: t(\alpha_1, \cdots, \alpha_n) \to [0, 1]\), which, for any \(r \in [0, 1]\), satisfies [58]

\[ P((z_1, \cdots, z_n) : t(\alpha_1, \cdots, \alpha_n) \leq r) \leq r. \quad (2.3) \]

That is, the strangeness measure function must define a space where Martin-Löf randomness test [49] can be applied. The confidence and reliability of the prediction is obtained by applying the approximate randomness test on strangeness measure [27].
The strangeness $\alpha_i$ of a particular example $x_i$ measures the uncertainty of that example with respect to its label and all other examples: the higher the measure, the higher the uncertainty. It in fact carries the discriminative information of that example. It can be used either for classification or feature selection, which will be shown in the later chapters.

In order to achieve clear confidence values for the prediction, the strangeness measure used depends on the type of base classifier used to construct the confidence machine. Several strangeness definitions are proposed recently. If the Support Vector Machine is used as the base classifier

$$y(x) = \text{sign}(\sum_{i=1}^{n} \alpha_i y_i K(x_i, x)), \quad (2.4)$$

the Lagrange multipliers $\alpha_i$ associated with each example can be used as strangeness measure [27]. The weights $w_i$ in kernel perception machine

$$f_w(x) = \text{sign}(w \cdot \phi(x)) = \text{sign}(\sum_{i=1}^{n} w_i K(x_i, x)) \quad (2.5)$$

are used as the strangeness measure in the work [58]. If $H$ is a fixed class of classification hypotheses, the “empirical log ratio” [22] is the strangeness of $x$ over $H$:

$$\hat{\ell}_\eta(x) = \frac{1}{\eta} \ln \left( \frac{\sum_{h, h(x)=+1} e^{-\eta \hat{\varepsilon}(h)}}{\sum_{h, h(x)=-1} e^{-\eta \hat{\varepsilon}(h)}} \right), \quad (2.6)$$

where $\hat{\varepsilon}(h)$ is the estimated error of hypothesis $h$ according to the training examples. Those proposed strangeness need complex learning strategy and high computational cost. There are several simpler definitions which do not require complex learning
procedure. If the example of class $j$ is sampled from a Gaussian model, the distance from example $x^i_j$ to the mean $\bar{x}^j$ can be defined as the strangeness:

$$\alpha_i = \| x^i_j - \bar{x}^j \|, \text{ where } \bar{x}^j = \frac{1}{N^j} \sum_k x^j_k.$$  \hfill (2.7)

Without any assumption about distribution $D$ of $z$, $k$-nearest neighbor classifier is widely used in [42,68,75] to define the strangeness measure if the examples are measurable in some metric space. Assume we have $C$ classes, for class $c = 1, \cdots, C$, let us denote the sorted sequence (in ascending order) of the distances of example $x^c_j$ from the other examples with the same class $c$ as $d^c_j$ and $d^c_{jl}$ will stand for the $l$th shortest distance in this sequence. Let $d^{-c}_j$ denote the sorted sequence of distances containing examples with class label different from $c$. For each example, the individual strangeness measure is assigned as:

$$\alpha_j = \frac{\sum_{l=1}^k d^c_{jl}}{\sum_{l=1}^k d^{-c}_{jl}}.$$ \hfill (2.8)

The measurement for strangeness is the ratio of the sum of the $k$ nearest distances from the same class to the sum of the $k$ nearest distances from all other classes. This definition of strangeness is very natural and straightforward. An example is considered strange if it is in the middle of examples labeled in a different way and is far from the examples labeled in the same way. The strangeness of an example increases when the distance from the example of the same class becomes bigger or when the distance from the other classes becomes smaller. The strangeness defined in Equation 2.8 is related to $k$-nearest neighbor classifier (kNN). However, for multi-class
classification, the definition in Equation 2.8 does not consider the frequency of each class in the neighborhood of the example, as does in kNN classifier. As the result, we modify the definition in Equation 2.8 and re-define the k-NN strangeness as:

$$\alpha_j = \frac{\sum_{l=1}^{k} d_{jl}}{\min_{n,n\neq c} \sum_{l=1}^{k} d_{jl}^n}.$$  \hspace{1cm} (2.9)

In the following section, we will discuss its properties and show how it is related to optimal decision boundary and the posterior $P(c_i|x_i)$.

### 2.3 k-Nearest Neighbor Strangeness

In this section we will study the properties of $k$-nearest neighbor strangeness, as defined in Equation 2.9, to show how it can be used to build the decision boundary between classes and how it is related to the discrimination ability of each example. The Cover-Hart theorem \[14\] proves that asymptotically the generalization error of one nearest neighbor classifier can exceed by at most twice the generalization error of the Bayes optimal classification rule. They also showed that the $k$-NN error approaches the Bayes error (with factor 1) if $k = O(\log n)$ \[14\]. The asymptotical generalization capability of $k$-NN classifier enables the $k$-NN strangeness has the similar properties.

On average the examples with $\alpha = const$ build the piecewise linear boundary between class $c_i$ and all other classes. The smaller the strangeness of the example is, the larger the distance of the example to the decision boundary and the higher discrimination ability of the example with respect to other examples and class labels. Asymptotically, the examples with $\alpha = 1$ will build the optimal boundary between
two classes. Those examples can be considered as samples from the optimal Bayes
classification boundary which serves as the ground truth if the data distribution
and prior are known. We demonstrate the comparison between those two decision
boundaries in synthetic experiments next.

We consider a two-class classification first. Let examples \((z_1, \cdots, z_n) = ((x_1, y_1),
\cdots, (x_n, y_n))\) be drawn independently from the same distribution over
\(Z = X^d \times Y\) where \(Y\) is the label space \(\{0, 1\}\). For each class \(c_i\), the data is generated independently
from a Gaussian distributions \(P(x|c_i) = N(x; \mu_i, \Sigma_i^{-1})\) and priors \(p_i = P(c_i), i = 0\ or 1\). Applying the Bayesian rule, the optimal boundary between two classes is a
quadratic function when two classes have different covariance matrix:

\[
f(x) = x^T(\Sigma_0^{-1} - \Sigma_1^{-1})x - 2(\mu_0^T \Sigma_0^{-1} - \mu_1^T \Sigma_1^{-1})x
\]

\[
+ (\mu_0^T \Sigma_0^{-1} \mu_0 - \mu_1^T \Sigma_1^{-1} \mu_1 - \log \frac{|\Sigma_1|}{|\Sigma_0|} - 2 \log \frac{p_0}{p_1})(2.10)
\]

If \(\Sigma_0 = \Sigma_1 = \Sigma\), the decision boundary becomes a linear hyperplane:

\[
f(x) = (\mu_0 - \mu_1)^T \Sigma^{-1}x - \frac{1}{2}(\mu_0^T \Sigma^{-1} \mu_0 - \mu_1^T \Sigma^{-1} \mu_1) + \log \frac{p_0}{p_1} = w^T \cdot x + b, \quad (2.11)
\]

where \(w = (\mu_0 - \mu_1)\Sigma^{-1}\) and \(b = \log \frac{p_0}{p_1} - \frac{1}{2}(\mu_0^T \Sigma^{-1} \mu_0 - \mu_1^T \Sigma^{-1} \mu_1)\). In order to demonstrate
the properties of strangeness, we compute the boundaries in two dimensional
space using the Gaussian example, \(kNN\) classifier and strangeness, respectively. Let
the means of two classes be \([0, 0]^T\) and \([5, 5]^T\), respectively. They have the same co-
variance matrix \(\Sigma = diag\{\sigma, \sigma\}\). Same number \(N\) of examples are sampled for both
classes, which means \(p_0 = p_1 = 0.5\). The following experimental results show the
optimal boundaries and boundaries constructed by \(kNN\) and \(kNN\) strangeness with
\( \alpha = 1 \), with respect to different \( \sigma, k \) and \( N \).

If two classes are well-separable, the boundaries constructed by kNN and kNN-strangeness are very close to each other, and both of them are close to the optimal boundary, even when both \( N \) and \( k \) are small. Given \( \sigma = 1 \), Fig. 2.2 shows the obtained boundaries. If \( k \) is small, those boundaries have little improvement even when \( N \) increases (comparing Fig. 2.2(c) with Fig. 2.2(a) and (b)). If both \( N \) and \( k \) increase, both boundaries converge to the optimal boundary, as shown in Fig. 2.2(d).

This is consistent with generalization ability of \( k \)-nearest neighbor classifier – kNN generalization error approaches the Bayes error with factor 1 if \( k = O(\log n) \).

Figure 2.2: The decision boundaries with different \( N \) and \( k \)
From the examples above, it seems that the definition of strangeness has no advantage over $k$-nearest neighbor classifier. In real application, however, the classes are rarely well-separable; and the data are often from very high dimensional space. We now examine the comparison with non-separable data set. Suppose $\sigma = 3$ and a small number of training examples is given, $N = 100$, both boundaries are not too close to the optimal boundary. The boundary constructed by strangeness is, however, much more smooth and closer to the optimal boundary. Fig. 2.3 shows the boundaries with different $k$. When $k$ is small, the strangeness smoothes many isolated regions created by $k$NN classifier. With respect to the optimal boundary, the boundary constructed by strangeness has much less variation than $k$NN, hence introduces a less complex functional space to build the inference boundary. Consequently it should have better generalization ability. If we consider the problem in a regularization framework, strangeness introduces a smooth penalty term, which is defined through the examples with the parameter $k$. Let $k = \log N$, Fig. 2.4 shows the boundaries with different $N$. The boundaries constructed highly depend on the training examples while they both converge to the optimal boundary as $N \to \infty$. For each $N$, the boundary created by strangeness is much more smooth.
Now let’s consider the test error. For each $N$, different training and testing sets are sampled in 100 trials. Fig. 2.5 shows the optimal Bayesian error, test errors of $k$NN and strangeness classifiers and the corresponding error standard deviation over the trials. The error of strangeness and its standard deviation are always lower than those of the corresponding $k$NN classifier, which is consistent with the conclusion from the comparison of the boundaries. The smooth term of the classification function reduces the test error and hence improves the generalization capability of the algorithm. Compared with $k$NN classifier, the strangeness not only predicts the label of the test examples, but also gives the “reliability” or “confidence” of the prediction – the value of the strangeness measure: the higher the measure, the higher the uncertainty of the prediction. Margin play the crucial role in modern machine learning research, which measure the classifiers confidence and generalization capability. Now we will study the relationship between strangeness and margin in two-class classification problem.

There are two natural ways of defining the margin of an instance with respect to a classifier. The more common type, sample-margin, measures the distance between the instance and the decision boundary induced by the classifier. Let the classification function be $f(x)$, the label space be $Y = \{-1, +1\}$ and the prediction rule
Figure 2.5: The test errors and their standard deviations with different $N$.

Figure 2.6: The relationship between strangeness and margin with different $N$.

be $y = \text{sign}(f(x))$, then the margin can be defined as $\phi(f(x), y) = yf(x)$. Using the same synthetic data above, Fig. 2.6 shows the relationship between the margin and the corresponding mean strangeness measure of $x_i$ with fixed $N$ training data examples, where $f(x)$ is the optimal boundary (the hyperplane in our case). In the experiments, the training data set is given; 5000 test examples are randomly sampled for each class. The mean strangeness of the test examples with same margin is then calculated. From the curve shown in Fig. 2.6, we can see that there exists a monotonically decreasing function between strangeness and the sample-margin. The function has the parameter with respect to training set $N$. It converges to some monotonic
decreasing function when $N \to \infty$. Consequently it can also be considered as one definition of “margin”. In fact, similar definition of margin is called hypothesis-margin, as used in AdaBoost [23]. The hypothesis-margin of an hypothesis with respect to an instance is the distance between the hypothesis and the closest hypothesis that assigns alternative label to the given instance. AdaBoost uses $L_1$ norm as the distance measure among hypotheses. The work proposed in [28] defines the hypothesis-margin using one nearest neighbor as following:

$$\phi(x) = \frac{1}{2k}(|x - \text{nearmiss}(x)| - |x - \text{nearhit}(x)|), \quad (2.12)$$

where $\text{nearhit}(x)$ and $\text{nearmiss}(x)$ denote the nearest point to $x$ with the same and different label, respectively. It is very close to our definition of strangeness in Equations 2.8 and 2.9. The similar hypothesis-margin using $k$ nearest neighbor can be defined as

$$\phi_j = \frac{1}{2} \left( \min_{n,n\neq i} \sum_{l=1}^{k} D_{jl}^n - \sum_{l=1}^{k} D_{jl}^i \right). \quad (2.13)$$

It has a linear relationship with sample-margin, as shown in Fig. 2.7. It is a measurement capturing the relevant properties of the data point. Since a good generalization can be guaranteed if many samples have large margins [76], the final classifier will have very good generalization capability if we can find a feature selection algorithm where the maximal strangeness measures of training examples are minimized.

The strangeness of test example $x$ is not only related to its margin, but also has the one-to-one mapping relationship with its posterior probability $P(c_i|x)$ as shown in Fig. 2.8(a) and (b). Given the likelihood functions $P(x|c_0), P(x|c_1)$ and the priors
Figure 2.7: The relationship between hypothesis-margin and sample-margin with different $N$

$p_0, p_1$, the posterior of example $x$ is computed using the Bayesian rule:

$$P(c_i|x) = \frac{P(x|c_i)p_i}{P(x|c_0)p_0 + P(x|c_1)p_1}$$ \hspace{1cm} (2.14)

Odds $\frac{P(c_0|x)}{P(c_1|x)}$ are often used in classification problems to determine the membership of test example $x$. Fig. 2.8 (c) and (d) shows the relationship between strangeness and the odds. Our strangeness definition can be considered as the local estimation of odds.

So far we only demonstrated the properties of $k$NN strangeness measure for Gaussian distribution in 2 dimensional space. Now we consider the properties of strangeness in high dimensional space, non-Gaussian distributions and multiple classes cases – which are more common in real world problem. Fig. 2.9 shows the optimal boundary, and the boundaries of strangeness and $k$NN, respectively, when two classes are mixture Gaussian distribution. Class 0 has the three modes with the means $\{[2, 2], [-1, 1], [5, 2]\}$ and covariance matrices $\{\text{diag}(1.5, 1.5), \text{diag}(1, 1), \text{diag}(1, 1)\}$, respectively. Class 1 has also three modes with the means $\{[1, -2], [-2, -1], [3, 0]\}$
and the same covariance matrices as class 0. Each mode has the same weight in both classes. For each class, \( N \) training examples are randomly drawn. The boundary constructed by strangeness is much more smooth than \( k \)-NN, which is consistent with 2D case. Running the experiments with different training data sets, Fig. 2.10 shows the average test errors and the corresponding error standard deviations with different size \( N \) of training sample. Obviously strangeness measure has always lower variance and better generalization capability. Its test error is always lower than ones of \( k \)-nearest neighbor classifier.

In general, the subjects are often represented by very high dimensional vector. It is
Figure 2.9: The boundaries constructed by kNN and strangeness with different $N$.

Figure 2.10: The test errors and their standard deviations with different $N$.

very important to study the properties of strangeness in high dimensional space. Consider two $d$-dimensional Gaussian distributions with means $[0, \cdots , 0]$ and $[5, \cdots , 5]$, respectively. Assume they have different covariance matrices such that the optimal classification boundary is no longer a hyperplane. The two covariance matrices are randomly generated in the following way. At first two $d \times d$ matrices $A$, $B$ are generated with all of the diagonal elements are 4 and 6, respectively. Their off-diagonal elements are randomly generated between $[1, -1]$. Then the covariance matrices are computed as $\Sigma_1 = AA^T$ and $\Sigma_2 = BB^T$, which guarantee $\Sigma_1$ and $\Sigma_2$ satisfying the
properties of covariance matrix. The distance of two instances is measured by Euclidean distance. In $d$ dimension space, each class has $N$ training examples randomly sampled from the distributions above. Another 10000 examples are randomly generated for testing. Fig. 2.11 shows the test errors of $k$NN and strangeness ($\alpha = 1$) in different dimensional spaces, $d$ from 2 to 100. It clearly shows the strangeness has between performance over $k$ nearest neighbor classifier no matter the dimensionality of the representation. If the number of training example number is fixed, both test errors increase a little bit as the dimension of the data is close the training example number (see Fig. 2.11 (a)). This is the realization of the “curve of dimensionality” [9]. Similar results can be obtained for high dimensional mixture Gaussian distribution.

![Figure 2.11: The test errors with respect to the dimensionality given the fixed $N$ training examples.](image)

Although both classifiers have close performance, the strangeness not only yields the “bare prediction” as $k$-nearest neighbor does, it also gives the confidence of the prediction. The strangeness measure has further monotonic relationship with margin, posterior and odds. If we can find a transformation or the metric space where the maximal strangeness measures of training examples are minimized, the final classifier
will have very good generalization ability – it can be considered as one maximal margin classification. Based on the strangeness measure, we can further threshold it with $\epsilon < 1$, which divides the space into three regions – two reliable prediction regions ($\alpha_j < \epsilon$) and one unreliable region ($\alpha_j \geq \epsilon$). The reliable regions yield the high confident prediction while the unreliable region tell us where the error occurs – identifying the locations of potential overfitting. In such a case, the system can allow the user to choose a special course of action for the examples falling in the unreliable region, as did in [22].

In a word, we have introduced the strangeness measure and discussed how it can be used as a classifier to build the decision boundary, measure the uncertainty of the prediction and its generalization ability. We also investigated its relationship to margin and posterior and its behavior of strangeness for high dimensional data and non-Gaussian distribution. In the following chapters, we will show how it can be used for feature selection, generic object recognition and open set recognition.
Chapter 3: Feature Detection, Representation and Recognition Approach

As noted in Chapter 1, one main purpose in this thesis is to establish a framework for part-based generic object recognition where the training images contain the objects as well as the clutter background. There are three main issues of a part-based object recognition approach: detect and/or choose object local features, represent them and model the geometrical configuration. We adopt object models consisting of a collection of small local features without considering their spatial relationship. Our learning model selects a “small set” of discriminative features, integrates the local discriminative evidence from them, and compensates for the repeatability of the feature detectors. This chapter briefly reviews previous work on local feature detectors and their associated descriptors. In particular, we describe SIFT features [53] and Affine Covariant Regions [62] that are used in this thesis. Since face is highly non-rigid object and can show big local distortion because of the facial expression, almost none of those feature detectors can successfully and repeatedly detect the same salient region from same person faces with different facial expression. To overcome this difficulty, we present a geometrical framework where candidates of local face patches are found.

At the end of this section we briefly outline our learning framework. The detail of our approach is relegated to Chapter 4 and 5.
3.1 Feature Detection

The central philosophy behind many feature detectors is to search for the “salient” regions and characterize them by a descriptor invariant to local deformation and illumination variation. One of the most popular feature detectors are Förstner corner [20] detector and Harris feature detector [32]. Förstner detected the corners by implicit least squares matching, using rectangular patches and Taylor expansion to re-express the accuracy in terms of the eigenvalues of the scatter matrix of the local gradients. Harris & Stephens [32] improved the localization performance by replacing the rectangular patches with Gaussian convolution window with a scale. Harris points are invariant to rotation and the support region is a fixed size neighborhood centered at the interest point.

Tuytelaars and Van Gool [81] construct small image regions depending on both corners and the intensity extreme. The intensity extreme are extracted based on the intensity profiles along rays emanating from the points. The nearby edges provide orientation and skew while the scale and stretch are given by the extreme of a 2D affine invariant function. Kadir and Brady [37] find regions that are salient over both location and scale, which are invariant to rotation and scale changes. In this work, the saliency is defined in terms of local signal complexity or unpredictability, which is measured with Shannon entropy of local attributes. For each point on the image a histogram $P(I)$ is made of the intensities in a circular region of radius (scale) $s$. The entropy $H(s)$ of the histogram is then calculated and the local maxima of $H(s)$ are candidate scales for the region. The saliency of each of these candidates is measured by $H \frac{dP}{ds}$ with appropriate normalization [37]. Matas et at [56] used a procedure to
construct a maximally stable affine co-variant regions. The areas are selected from an intensity images using watershed image segmentation. The proposed regions are the connected components of pixels which are all brighter or darker than all pixels on the region’s contour and are bounded by cycles of edge pixels.

To obtain invariance to scale changes, interest points can be extracted in the scale space of the image [50]. Mikolajczyk & Schmid [60] improved Harris Corner detector by introducing so called Harris-Laplace points, which are invariant to rotation and scale changes. The points are detected by a scale adapted Harris measure and selected in scale space by the Laplacian operator. The selected scale determines the size of the support region. But in case of significant transformation, the feature detector has to be adapted to the transformation. Mikolajczyk & Schmid [38,62] further introduced the Harris-Affine point detector. The detected features are invariant to affine image transformation and constructed by elliptical shape adaptation. The outline of the detector is represented in the following:

- The spatial localization of an interest point for a given scale and shape is determined by the affine-adapted Harris detector. Multi-scale Harris detector is used to get the initial localizations and scales. An iterative procedure is applied to modify the position, scale and the shape of point neighborhood, which converge to a stable point invariant to affine transformations.

- The integration scale is selected at the extremum over the scale of normalized derivatives.

- The derivation scale is selected at the maximum of normalized isotropy

- The shape adaption matrix normalizes the point neighborhood.
Fig. 3.1 shows the affine covariant regions detected in the images, which are used for the weakly supervised generic object recognition in this thesis. According to the report in [64], the affine covariant region detector has good performance over the viewpoint change, local deformation and illumination. It provides enough features for matching and recognition and robust to clutter and partial occlusion.

Another powerful feature detector is Lowe Detector (DoG points) [53], which is invariant to rotation and scale changes. It is implemented efficiently by constructing a Gaussian pyramid and searching for local peaks in the scale space of difference-of-Gaussian (DoG). Candidate feature points are localized to sub-pixel accuracy and selected based on measures of their stability. The selected scale determines the size of the support region. Fig. 3.2 shows the extreme of DoG regions detected. Compared with affine covariant region detector, Lowe’s detector detects more regions in the images, especially in the images with complex texture. We use it for the location recognition task, where no much texture is present.
3.2 Feature Description

The detected features, local regions associated with, are further described by a descriptor, which is robust to the local deformation, rotation/affine transformation and illumination. The invariance properties are essential in order to be able to handle variations due to viewpoint, deformation and lighting conditions. The simplest descriptor is a the image pixels. But the high dimension of such a descriptor increases the computational complexity of recognition, and it is sensitive to many variations.

Several more advanced feature descriptors were proposed in the past few years. Johnson and Hebert [36] used the distribution based descriptor for 3D object recognition. Zabih and Woodfill [92] developed the non-parametric transform, rank transform and census transform, to describe the local regions. It is highly robust to illumination change. Spatial-frequency techniques are also often used to describe the image content. They are include Fourier transform, Gabor filters and wavelets. In order to deal with rotation, Freeman and Adelson [21] developed differential steerable descriptors for image content. A stable estimation of the derivatives is obtained by convolution with Gaussian derivatives. Complex filters are used in [8, 25].
to achieve the rotation invariance. In those works, the complex filter bank was used from the family $K(x, y, \theta) = f(x, y) \exp(i\theta)$. Generalized moment invariants are also popular to describe the multi-spectral nature of the data, which is defined as: $M_{pq}^a = \int \int x^p y^q [I(x, y)]^a dx dy$ with order $p + q$ and degree $a$. These moments are independent and can be easily computed for any order and degree.

![Image of SIFT Feature Descriptor](image)

Figure 3.3: SIFT Feature Descriptor.

David Lowe recently presented a very robust descriptor so called Scale Invariant Feature Transforms (SIFT) in [53], as shown in Fig. 3.3. Once the location and scale have been assigned to candidate features, the dominant orientation is computed by determining the peaks in the orientation histogram of its local neighborhood weighted by the gradient magnitude. The descriptor is then formed by computing local orientation histograms (with 8 bin resolution) for each element of a $4 \times 4$ grid overlayed over $16 \times 16$ neighborhood of the point. This yields a 128 dimensional feature vector which is normalized to unit length in order to reduce the sensitivity to image contrast and brightness changes. This descriptor provides robustness against localization errors and small geometric distortions, and has good repeatability across variations of scale and pose. Mikolajczyk & Schmid [61] evaluated those descriptors based on ROC curves of detection rate under rotation, scale change, affine transformation and illumination changes. The results show that SIFT descriptors obtain better results
than others except for light changes. As a result, we adopt SIFT feature descriptor to model the appearance of local region in our recognition framework.

### 3.3 Part-based Face Representation

The holistical appearance approach was widely used in face recognition domain [2, 47, 51, 80], where the faces are encoded via principal component analysis (PCA) and represented as eigenimage. Although Leonardis et al [3] further proposed a robust eigenimage recognition algorithm, which considered the partial occlusion, it is still hard to deal with occlusion, local facial deformation and disguise. It has poor class generalization ability. In addition, it is quite hard to be met in practice.

To compensate for pose changes and possibly for limited occlusion and distortions, Elastic Graph Matching (EGM) [45, 90], representative of part-based face recognition, modelled face in a scale space and flexible and non-rigid string geometry that connects the nodes across the face. The parts or nodes define a set of landmark points, which are coarsely represented using Gabor jets. Training each face component classifier separately, Ivanov et al. [34] have explored simple strategies for combining SVM classifiers for the components of the faces. It showed a marked improvement compared to holistical face classifiers. One major drawback of above representation is that all local patches are extracted in the same scale, which is not enough to represent the global and local information at the same time and is not adaptive for different subjects. Mikolajczyk et al [63] have proposed a framework based on a probabilistic assembly of robust human part detectors. It requires precise manual annotation for the parts and extensive training to approximate a Bayesian model. Reliable location for the face and its parts is, however, a hard problem. It is basically for two class
As we mentioned before, face is highly non-rigid and can show large local distortion due to facial expression. Almost none of the existing feature detectors can successfully and repeatedly detect the same local region across different images from same person. To overcome this difficulty, we use a face template, golden ratio template, to represent the face with multiple local patches spanning a multi-resolution grid to address the above issues and expand the existing paradigms.

Golden ratio face templates are built on biological principles and was originally proposed by Sinha [78]. Anderson and McOwan [4] modified it and successfully applied for face detection in real time. The template provides a rough spatial map of the facial features, and is informative about which facial regions are important for recognition. Figure 3.4 shows the local regions in the template. Instead of computing the ratio of the average gray-values from different regions to represent the face as did in [4,78], we treat each region center as a feature point and extract the face components with different scales and different bandwidth channels.

Given the positions of eyes, the center $o_i$ and minimal width/height $r_i$ of each
region in the face image is determined by the size of each region in the template. Let \( o_i \) be the center of the feature we want to extract, and \( r_i \) be the initial scale of each feature region. In order to encode the local and global information of the face, multiple face components are extracted with different scale at each position. The \( k \)-th component of region \( i \) has the scale \( r_{ik} = s^{k-1}r_i, \) \( s = \sqrt{2} \) in our experiments. Furthermore, a Gaussian pyramid [53] is built by blurring the original image. The same number of patches are extracted at each level of the pyramid. Since each region in the template is considered independently, we call those local patches as 1st order patches. Given that the golden ratio template consisting of 16 regions, and given the scale level \( N_s \) and the blurring level \( N_b \), there are \( 16N_sN_b \) 1st order local patches extracted from one face image. We further extract high local patches, which is motivated by Barlow’s suspicious coincidences [6]: two candidate feature \( A \) and \( B \) should be encoded together if the joint appearance probability \( P(A, B) \gg P(A)P(B) \). To reduce the complexity of the model learning, we only extract 2nd order local patches from two local regions which are neighbors to each other in the template. The size of each patch is represented by an ellipse with center \( \overrightarrow{x} \) and parameters \( a \) and \( b \). Given two neighborhood regions \( i \) and \( j \) with scale \( r_{ik}, r_{jk} \) and centers \( \overrightarrow{x}_i, \overrightarrow{x}_j \), the 2nd local patch is extracted with \( \overrightarrow{x} = (\overrightarrow{x}_i + \overrightarrow{x}_j)/2, \) \( a = (\|\overrightarrow{x}_i - \overrightarrow{x}_j\| + (r_{ik} + r_{jk}))/2 \) and \( b = \max(r_{ik}, r_{jk}) \). There are in total \( 27N_sN_b \) 2nd order local patches extracted. Figure 3.5(a) and (b) shows the 1st and 2nd order local patches with the initial scale, respectively.

The next step is to compute a descriptor for each local patch that is highly distinctive yet is as invariant as possible to the variations, such as illumination and
deformation, e.g., facial expression. We use the SIFT descriptor to represent each local region, as described in Section 3.2. For efficiency, the gradients are pre-computed for all levels of the blur pyramid. In the end, each face is represented by $43 N_s N_b$ local features with SIFT descriptor.

### 3.4 Part-based Object Recognition Approach

As discussed in the introduction, the part-based representations have been shown favorable in the context of real-world recognition task to handle the large variations as well as clutter background. The simplest method for part-based recognition is so called voting approach. In the voting scheme, the recognition is determined by the number of matched features from the query image to model views in training database. Lowe [53] developed a robust matching procedure to discard large number of false matches. The works in [19, 38, 60–62, 64] also used voting for object matching. Voting scheme, however, only works well for rigid object matching. It does not model the uncertainty of local feature. Neither does it provide the confidence of the
recognition. Furthermore, it requires a threshold to determine the matching, which is difficult to get through the limited training data since different local features have different appearance variations. In order to capture the uncertainty of descriptor and spatial relationship, probabilistic approach is widely used where the objects are modelled as random constellations of parts. The works in [17, 18, 33] modelled the feature appearance and shape together in a Bayesian framework. In those works, the appearance and shape were assumed to be the mixture Gaussian distribution. In order to reduce the complexity of model learning, FeiFei et al [17] modelled the prior of the parameters with a Normal-Wishart distribution. The proposed Bayesian framework, however, only has the object detection capability, which means it can only distinguish object category from background and has no discriminant ability among the object categories. Due to the complexity of the model, it is difficult to generalize for individual object recognition, which need large number of parts to discriminate the objects. More comments are given in Chapter 5. Considering the local feature appearance only, the works in [1, 5, 15, 24, 77, 87, 88] recognized the objects with different machine learning models from local features. They either need pre-segmented/aligned training data or are sensitive to percentage of the irrelevant features in the image.

The motivation of our approach is shown in Fig. 3.6. We assume that each object/category be represented by a collect of discriminant components, which are common among the same object subject/category while discriminative from other objects/categories. In case of the presence of the clutter background, strangeness-based feature selection algorithm is applied to discard the irrelevant features and select the discriminative ones, which further build the parts of the objects and yield the local
evidence provided by individual features for recognition. Regarding part-based face recognition, all extracted features are relevant and directly used to build the face components because of the biological principles of the face template. If the training images are pre-segmented and the features are extracted by scanning the images pixel by pixel, as did in [24, 87], our approach is still applicable. Strangeness-based feature selection algorithm can be applied first to reduce size of feature pool thus reducing the computational and model complexity of the learning procedure. Finally, general boosting strategy is used to integrate the local evidence, characterized by strangeness, and fuse them into final strong classifier. Compared with other part-based object recognition approach, it has no complex parametric model assumption and parameter estimation. It is applicable to both individual object and object category recognition, which is demonstrated in the later chapters.

Figure 3.6: The motivation of our approach.

The outline of the approach is shown as following:

1. Local features are extracted from the training images.
Figure 3.7: Samples images of categories database contain faces, airplanes, cars, motorbikes and the background.

2. The discriminative features are selected, if necessary, using strangeness measure.

3. The parts of object are built.

4. Object/category is represented by parts associated with the feature instances.

5. Strangeness-based weak classifier is employed in boosting framework, where coefficients and thresholds are obtained through validation.

6. Hierarchical classifier is applied for weakly supervised learning.

7. Test the query image with the learned the classifier.
For convenience and to show the difficulty of the problem, representative images from the object category database are shown here in Fig. 3.7 containing objects on varying background. In Chapter 4 we will show how strangeness is used to deal with the clutter background. In Chapter 5 we will present our learning algorithm and the experimental evaluation with different databases.
Chapter 4: Strangeness-based Feature Selection

Chapter 2 presents the definition of the strangeness measure, its properties and generalization capability. The strangeness measure not only predicts the label, but also gives the relevance between samples or feature instances and classes. In this chapter, we will present how such measurement is used as a criterion to do feature selection. Using the same strangeness definition, different feature selection algorithms are presented to do variable selection and feature instance selection, respectively.

4.1 Related Work

In many supervised learning tasks, the input data is often represented by a very large number of features. Even state-of-art learning algorithms cannot overcome the presence of large number of weakly relevant or irrelevant features. This is attributed to the “curse of dimensionality”, or the fact that irrelevant features decrease the signal-to-noise ratio and confuse the learning algorithm. In addition, many learning algorithms become computationally intractable when the dimension is high. On the other hand, once a very good set of features is obtained, even the very basic and simple classifiers can achieve high performance. As a results, feature selection is crucial for efficient learning. Furthermore, there are many other potential benefits of feature selection: facilitating data visualization and data understanding, reducing the measurement and storage requirements, reducing the complexity of learned models, and defying the curse of dimensionality to improve prediction performance.
Various aspects of feature selection have been studied. One of the key aspects in feature selection is to measure the *goodness* of a feature subset to determine the optimal one. Different feature selection algorithms can be broadly divided into two categories: *filters* and *wrappers*. The filter approaches evaluate the relevance of each feature (subset) using the data set alone, regardless of the subsequent learning phase. Because performing exhaustive search is usually intractable, a variety of search heuristics are applied. *Relief* [74] and information theoretic methods [24, 44, 66, 85] are the representatives of this class. *Relief* assigns feature weights based on the consistency of the feature value in the $k$ nearest neighbors of every data point. The philosophy behind the information theoretic methods is that the mutual information between a relevant feature and the class labels should be high [7]. Nonparametric methods can be used to compute the mutual information involving the continuous features [44]. A feature can be regarded as irrelevant if it is conditionally independent of the class labels given other features. The concept of Markov blanket is used to formalize this notion of irrelevancy in [40].

On the other hand, the wrapper approach [5, 33, 39, 87] invokes the learning algorithm to evaluate the quality of each feature (subset). A learning phase, such as boosting [5, 87], Bayesian approach [18], decision tree [66] is run on a feature subset and the feature relevancy is assessed by the estimation of the classification accuracy. Wrappers are usually more computationally demanding, but can be superior in accuracy when compared with filters. Both approaches involve combinatorial searches through the space of possible feature subsets with different types of heuristics.

Given the training features $F = (F_1, \cdots, F_N)$, where $F_i$ is a point in $\mathbb{R}^d$, there are two different feature selection directions: one is to select the optimal subspace
along the column direction of the feature matrix \( F \) - \textit{variable selection}; the other one is to select the optimal sub-instance along the row direction of \( F \) - \textit{feature instance selection}. The first one is widely researched in machine learning field, where one Assumes each instance of \( F \) has some contribution for classification and tries to find the optimal subspace and compact representation \([7, 39, 44, 74, 85]\). The second one is widely investigated in computer vision community where the objects are represented by local patches (features) (corresponding to feature instances in \( F \)) and the feature selection algorithms try to find the most discriminative and relevant patches to represent the objects \([18, 24, 33, 66, 87]\). Finding the optimal subspace is also widely used in computer vision field, such as Principal Component Analysis (PCA), Fisher Linear Discrimination (FLD) analysis and Independent Component Analysis (ICA) for holistic object recognition. Few works have investigated the feature selection in both directions of the feature matrix \( F \). In the following sections, we will present strangeness based feature selection algorithms in either direction and an empirical evidence on the performance of these algorithms.

### 4.2 Variable Selection

In this scenario, the training set is often given with labelled fixed-length feature vectors. Each feature is described as an assignment of values \( f = (f_1, \ldots, f_d)^T \) with \( d \) dimension to a set of features \( F = (F_1, \ldots, F_N) \) and one of \( l \) possible classes \( Y_1, \ldots, Y_l \). Feature selection is the task to try to find the optimal subset from \( F \) which has equal or as close as possible classification performance to the one using the whole feature set. It is closely related to more general problems of dimensionality
reduction and efficient data representation, while it provides a much simpler learning approach as it does not require the evaluation of new complex functions of the irrelevant features.

Mutual information between features and class labels is commonly used in this area as the evaluation function. Based on Fano’s bound, the Bayes error of $M$-class classification problem is lower bounded by the mutual information plus some constant:

$$p_e \geq \frac{H(X|Y) - H(p_e)}{\log(M - 1)} \geq \frac{H(Y) - I(X;Y) - 1}{\log(M)}.$$  \hspace{1cm} (4.1)

Thus the lower bound on error probability is minimized when the mutual information between $X$ and $Y$ is maximized. Alternatively, finding such features subset achieves the lowest possible bound to the error of a classifier.

Parzen window approach is often used to estimate the mutual information \cite{44}. However, it is hard and unreliable to be evaluated for high dimensional data with limited number of samples. To overcome this problem, the marginal probability and marginal diversity at each feature dimension is estimated in \cite{85}. The feature selection is based on the maximal marginal diversity criterion. This criterion is optimal for one dimensional features and optimal for multiple dimensions if the mutual information between features is not affected by the knowledge of the class label. This assumption may not be held for many feature representations. Vasconcelos further decomposed $F$ into several mutually exclusive feature subsets and used same criterion to do the feature selection \cite{86}.

Based on the performance of each feature with respect to the classification, the features can be divided into three categories: strong relevance, weak relevance and
irrelevance. Using this category information, one explicit feature redundancy analysis was proposed in [91] based on the feature’s Markov blanket [40]. Koller et.al [40] obtained an optimal subset by a backward elimination procedure, so called *Markov blanket filtering*. Let $G$ be the current set of features, a feature is redundant and hence should be removed from $G$ if and only if it is weakly relevant and has a Markov blanket $M_i$ within $G$ [91]. However measuring the conditional dependency for high dimensional data points is much hard in real practice.

**Definition 4.1. (Markov blanket)** Given a feature $F_i$, let $M_i \in F(F_i \notin M_i)$, $M_i$ is said to be a Markov blanket for $F_i$ if and only if

$$P(F - M_i - F_i, C|F_i, M_i) = P(F - M_i - F_i, C|M_i)$$

A good generalization can be guaranteed if many sample points have small strangeness, that is, the large margin (see Chapter 2). Instead of estimating mutual information, we introduce an evaluation function which assigns a score to sets of features according to the strangeness they induce. Similar to Relief algorithm, we first introduce a weight vector $\omega \in \mathbb{R}^d$ over the feature set $F$. Then the strangeness is re-formulated as the function of the weight vector $\omega$:

$$\alpha_j^\omega = \frac{\sum_{l=1}^k d(\omega)_{jl}^c}{\sum_{l=1}^k d(\omega)_{jl}^{-c}}, \quad (4.2)$$

or

$$\alpha_j^\omega = \frac{\sum_{l=1}^k d(\omega)_{jl}^c}{\min_{n,n \neq c} \sum_{l=1}^k d(\omega)_{jl}^n}, \quad (4.3)$$

where $d(\omega)_{jl} = \|F_j - F_l\|_\omega = \sqrt{\sum_k (F_{jk} - F_{lk})^2 \omega_k^2}$. Since $\alpha_j^{\lambda\omega} = \alpha_j^\omega$, it is not necessary
to normalize $\omega$. Given the training set $F$ and the weight vector $\omega$, we define the evaluation function as

$$
\Phi(\omega) = \sum_{F_j \in F} \alpha_j^\omega. \quad (4.4)
$$

Now we take our feature selection approach. First we find the weight vector $\omega$ which minimizes the evaluation function $\Phi(\omega)$ as defined in equation 4.4, and then use a threshold on $\omega$ to get the selected feature set as similar to Relieff algorithm. Since $\Phi(\omega)$ is smooth almost everywhere, gradient descent method is used to minimize it.

The gradient of $\Phi(\omega)$ with respect to the weight $\omega$ is:

$$(\nabla \Phi(\omega))_i = \frac{\partial \Phi(\omega)}{\partial \omega_i} = \sum_{F_j} \frac{\partial \alpha_j^\omega}{\partial \omega_i} = \sum_{F_j} \frac{\partial (\varphi_1(\omega)/\varphi_2(\omega))}{\partial \omega_i} \quad (4.5)$$

where $\varphi_1(\omega) = \sum_{l=1}^k D(\omega)_{jl}^c$, $\varphi_2(\omega) = \sum_{l=1}^k D(\omega)_{jl}^r$ or $\varphi_2(\omega) = \min_{n,n \neq c} \sum_{l=1}^k D(\omega)_{jn}^r$ with different definition of strangeness in equation 4.2 and 4.3. Take the partial derivative of $\varphi_1(\omega)$ and $\varphi_2(\omega)$ with respect to $\omega_i$ and we get:

$$
\zeta_1(\omega)_{\omega_i} = \frac{\partial \varphi_1(\omega)}{\partial \omega_i} = \omega_i \sum_{l=1}^k \frac{(F_{ji}^c - F_{li}^c)^2}{D(\omega)_{jl}^c} \quad (4.6)
$$
and

\[ \zeta_2(\omega) \omega_i = \frac{\partial \varphi_1(\omega)}{\partial \omega_i} = \omega_i \sum_{l=1}^{k} \frac{(F_{ji}^c - F_{li}^c)^2}{D(\omega)_{jl}^c} \text{ or} \]

\[ \zeta_2(\omega) \omega_i = \frac{\partial \varphi_1(\omega)}{\partial \omega_i} = \omega_i \sum_{l=1}^{k} \frac{(F_{ji}^c - F_{li}^n)^2}{\min_{n,n \neq c} \sum_{l=1}^{k} D(\omega)_{jl}^n} \]

Then the gradient of \( \Phi(\omega) \) is then computed as:

\[
(\nabla \Phi(\omega))_i = \sum_{F_j} \left( \frac{\zeta_1(\omega) \varphi_2(\omega) - \zeta_2(\omega) \varphi_1(\omega)}{\varphi_2^2(\omega)} \right) \omega_i
\]

Using a stochastic gradient descend over evaluation function \( \Phi(\omega) \), an iterative strangeness feature selection algorithm is proposed in Algorithm 1, where \( \rho \) is the learning rate. For \( N \) training instance in \( d \) dimensional space, the computational complexity of Algorithm 1 is \( O(NTd) \). The most complex operation is selection of \( k \) nearest neighbors to compute the strangeness. \( k - d \) tree can be applied to extract \( k \) nearest neighbors and speed up the algorithm, where the computational complexity becomes \( O(dN\log N) \).

Algorithm 1. Iterative Strangeness Feature Selection

1. Initialize \( \omega = (1, \cdots, 1) \).

2. For \( t = 1, \cdots, T \)
   - Pick randomly an instance \( F_j \) from training data \( F \).
   - Calculate the strangeness of \( F_j \) with respect to \( \omega \).
For $i = 1, \ldots, d$, calculate

$$\Delta_i = \left( \frac{\zeta_1(\omega) \varphi_2(\omega) - \zeta_2(\omega) \varphi_1(\omega)}{\varphi_2^2(\omega)} \right) \omega_i$$

- Update $\omega = \omega - \rho \Delta$

3. The selected feature set is $\{i|\omega_i > \tau\}$, where $\tau$ is a threshold.

where $\varphi_1(\omega), \varphi_2(\omega), \zeta_1(\omega)\omega_i$ and $\zeta_2(\omega)\omega_i$ are defined above and $\rho$ is the learning rate.

Relief is a popular feature selection algorithm related k-NN classifier (see Algorithm 2), which was demonstrated to be very efficient for estimating feature relevant quality. Under some assumptions, the expected weight is large for relevant features and small for irrelevant ones. It can deal with multi-class problems, noise and missing data [74]. Comparing with Relief algorithm, the updating rule in our algorithm is very similar to the one in Relief. However, the good generalization capability of strangeness (see Chapter 2) enables the samples to have large margin with the selected features. Thus more efficient learning algorithm and high performance can be obtained. In addition, Relief does not re-evaluate the distances according to the weight vector $\omega$ while our algorithm does. In term of computational complexity, two algorithms are equivalent.

**Algorithm 2. Relief algorithm** [74]

1. Initialize $\omega = (0, \cdots, 0)$.

2. For $t = 1, \cdots, T$

   - Pick randomly an instance $F_t$ from training data $F$ and let $c_t = \text{class}(F_t)$. 
• Find $k$ nearest hits $H_j$.

• Find $k$ nearest misses $M_j(c)$ for each class $c \neq c_t$.

• For $i = 1, \cdots, d$, calculate

$$
\omega_i = \omega_i - \frac{1}{mk} \left( \sum_{j=1}^{k} \phi(i, F_t, H_j) + \sum_{c \neq c_t} \left( \frac{P(c)}{1 - P(c_t)} \sum_{j=1}^{k} \phi(i, F_t, M_j(c)) \right) \right)
$$

3. The selected feature set is $\{i|\omega_i > \tau\}$, where $\tau$ is a threshold.

### 4.3 Feature Instance Selection

In order to deal with large variation of object appearance, due to occlusions, pose variations, deformation and size, many appearance-based approaches to object recognition characterize the objects by the local image features (patches) [18, 24, 33, 66, 87]. Each image is represented by $M_i$ features $\{g_j\}$ in $d$ dimensional space. The image patches are extracted by pixel-wise filtering [24, 66, 87] or by feature detectors [18, 33, 53]. To improve the performance of the final classifier, several methods have been proposed to select the feature instance. Viola and Jones [87] used an AdaBoost trained classifier to select the discriminative rectangular local features. Mahamud and Herbert [54] found discriminative object parts and develop an optimal distance measure for nearest neighbor search. Dorko and Schmid [15] developed one classifier for each object part, which was obtained by clustering the descriptor of local features. The object parts were selected according to the performance of the classifiers on the validation data set using the likelihood ratio and mutual information. Similar architecture has been applied for people detection [65]. However, in those works,
the training images are pre-segmented and complexity \textit{wrapper} approaches are used. Many irreverent background pixels are eliminated by hand. The works in \cite{65, 87} even need alignment.

Regarding the weakly supervised learning, only the label of the whole image is given while the location of the object in the image is unknown. The goal is to obtain object representation in terms of parts without further segmentation. In such scenario, each image has different number of local features detected as well as the large percentage of irrelevant features due to the clutter background. At the same time, the correspondence of local features is unknown between the images of the same classes. Because of the repeatability of the feature detectors, the same local feature cannot guarantee to be detected in all images of the same class. As a result, most \textit{filter} feature selection approaches can not work appropriately in such case. \textit{Wrapper} approaches with certain classifiers are often used, such as Support Vector Machines and Gaussian Mixture Model in \cite{15}, Bayesian framework in \cite{17, 18, 33} and linear Perceptron classifier in \cite{82}, as discussed above. Those approaches either need pre-segmented training data or complex learning method to construct the model with EM-type searching. Because of computational complexity of learning, the Bayesian framework approach in \cite{17,18,33} only works with the small number of parts ($P \leq 10$) and can distinguish the object from the background but has not enough discriminative power to separate one object from other objects.

Clustering is another popular approach to deal with irrelevant features from cluttered background. All features in training data set are clustered in certain number of clusters and the each training image is represented by the histogram of the cluster centers. With such representation, many standard feature selection algorithms can
be directly applied. Bishop [82] uses automatic relevance determination to discard the clusters belonging to the background. However, the number of clusters and clustering algorithm have great influence on the performance and generalization ability of the final classifier. Since the features from the background are assumed to be distributed uniformly in the descriptor space, large number of irrelevant features may have large number of clusters and overwhelm the relevant features in the clustering algorithm. In the end, totally different features coming from different objects may share the same cluster. In order to deal with the background clutter, we propose a simple and efficient algorithm to discard irrelevant features and select discriminative features for later learning stage, as shown in Algorithm 3. The algorithm is based on the strangeness measure used to evaluate the relevance between each local feature and the class label of the whole image.

Algorithm 3. Strangeness Feature Instance Selection

1. Given local features \( \{g_i\} \) in \( \mathbb{R}^d \) and class labels.

2. Compute the strangeness of each feature \( g_i \) based on Equation 2.9.

3. Initialize the threshold of strangeness \( \gamma \).

4. for \( t = 1, 2, ..., T \)
   - Select the features \( \{g_k\} \) with the strangeness \( \alpha_k \geq \gamma \).
   - Discard \( \{g_k\} \) and update the strangeness of remaining features.
   - If the strangeness of all features is less than \( \gamma \), terminate.

5. end
Strangeness Instance Feature Selection algorithm 3 is an iterative backward elimination method. The algorithm repeatedly iterates over the feature set and updates the set of chosen features. There is one threshold in the algorithm $\gamma$, which determines the features to be eliminated in each iteration and controls the largest strangeness, that is, the minimal margin, of the chosen features in the end. In each iteration, the strangeness of each feature is recomputed if necessary. The algorithm can be applied very efficiently if suitable data structure are used, because only small portion of strangeness need updating in each iteration. Compared with other feature selection algorithms, Algorithm 3 not only has the advantage of filter approaches – evaluating the relevance of feature and simple, but also have the properties of wrapper approaches – related to the predictor generalization performance. In fact Algorithm 3 is one approximation of the Markov blanket (see definition 4.1) of the irreverent features. Because we don’t model the geometric relationship between the feature instance, we can assume the features are independent. For irrelevant feature $g_k$, we denote it $k$ neighbors denoted as set $N_k = \{g_j\}_k$ and rest features $M_k = \{g_i\} - N_k - g_k$. According to the definition of Markov blanket, we have

$$P(M_k, C|g_k, N_k) = \frac{P(M_k, C, g_k, N_k)}{P(g_k, N_k)} = \frac{P(M_k, C, g_k|N_k)P(N_k)}{P(g_k, N_k)}.$$  \hspace{1cm} (4.9)$$

Since the feature $g_k$ has the strangeness $\alpha_k \geq \gamma$, it is irrelevant to the class labels, which means $P(M_k, C, g_k|N_k) \approx P(M_k, C|N_k)P(g_k|N_k)$. As a result, we can get

$$P(M_k, C|g_k, N_k) \approx \frac{P(M_k, C|N_k)P(g_k|N_k)P(N_k)}{P(g_k, N_k)} \approx \frac{P(M_k, C|N_k)P(g_k, N_k)}{P(g_k, N_k)} = P(M_k, C|N_k).$$  \hspace{1cm} (4.10)$$
Therefore the neighbors of irrelevant feature $g_k$ is its approximated Markov blanket during the feature selection.

4.4 Experiments and Evaluation

4.4.1 Variable Selection

To demonstrate the quality of the strangeness based evaluation function and the ability of Iterative Strangeness Feature Selection algorithm (Algorithm 1), we use a small synthetic toy problem. The problem consisted of 200 samples points with 10 real valued features. The target concept is the sign function over the product of the first 3 features. Hence, the first 3 feature are relevant while the others are irrelevant. This task is a special case of parity function learning and is hard for many feature variable selection algorithms [31]. We applied both Algorithm 1 and Relief on the samples with one nearest neighbor to do feature selection. Nearest neighbor classifier is further used to test the effect of feature selection on 500 test points. Fig. 4.1 shows the accuracy achieved versus the number of features selected. Algorithm 1 significantly outperformed Relief algorithm. Since feature variable selection is unnecessary in our following recognition tasks, we are not going to discuss its performance deeply in this thesis.

4.4.2 Feature Instance Selection

In this section, we first demonstrate the behavior and performance of Strangeness Instance Feature Selection (Algorithm 3) on a small synthetic two-class classification problem. Then we test it on a task of feature selection for discriminating between
Figure 4.1: The accuracy achieved when using the features chosen by both algorithms.

Consider two classes with different kinds of features sampled from different distributions. As shown in Fig. 4.2(a), the first class in red has two kinds of features sampled from two distributions: Gaussian distribution $D_1$ with mean $[0, 0]^T$ and standard deviation $\sigma = 2$, and uniform distribution $D_0$ over region $(3.5, 8.5) \times (-8.5, -3.5)$; the second class in green also has two kinds of features sampled from two distributions: Gaussian distribution $D_2$ with mean $[3, 5]^T$ and standard deviation $\sigma = 2$, and uniform distribution $D_0$ over region $(3.5, 8.5) \times (-8.5, -3.5)$. Two different classes have the features sampled from the same distribution $D_0$, which is very common in weakly supervised object recognition task. If no feature selection is applied, it is very hard to find a learning algorithm to get the optimal decision boundary and achieve high classification performance. For each distribution in each class, 300 points are randomly sampled as the training data set. Fig. 4.2(b) shows the selected features after Algorithm 3 is applied, where $\gamma = 0.8$. As we can see from the figures, most informative feature points are kept and most features with low discriminative ability are discarded. Only very small number features is chosen from $D_0$. Good decision
and high performance can be now achieved if the learning algorithm is applied on the selected features. Fig. 4.3 shows the results for another two synthetic data set.

![Graph showing original and selected feature instances.](image)

(a) Original feature instances. (b) Selected feature instances.

Figure 4.2: The features in training data set and the results after feature selection.

Now we apply Algorithm 3 on location recognition for robot navigation task. The video sequence is partitioned into 18 different locations, each of which is represented by different number of views. Each individual view is represented by scale-invariant (SIFT) keypoints described in [53]. More detail information can be found in [41,46]. Fig. 4.4 shows the detected and selected feature in one frame. The features belonging to posters have good discrimination capability. Fig. 4.5 shows the most likely selected features for each location and Fig. 4.6 shows the examples of non-discriminative features discarded by Algorithm 3, which are common in different location. From the figures, we can see the selected features have much meaningful information and characterize each location locally and globally. Note that some features in Fig. 4.5 have very large scales. The recognition results in Chapter 5 further demonstrate the feature selection are very effective.

We further applied Algorithm 3 on object category database used in [17,18]. We choose 4 categories: faces, motorbikes, airplanes and cars, as shown in Fig. 3.7.
For each class, 30 images are randomly chosen as the training gallery as well as the background. The features are detected by affine covariant regions detector and represented by SIFT descriptor (see chapter 3). Fig. 4.7 shows the original features detected and the features chosen by our algorithm on two example images. Note that most of the selected features are on the objects while most background features are discarded. The feature selection is very effective, which can make the learning stage more feasible and efficient, as demonstrated in Chapter 5.
Figure 4.4: a) The total of 480 features detected and b) 50 discriminative features selected.

4.5 Summary

A strangeness-based criterion for measuring the quality of a set of features has been presented. Using this criterion we derived algorithms that perform feature selection by searching for the subset that minimizes it. It is applicable for both feature variable selection and feature instance selection. The variable selection algorithm outperform the Relief from a simple toy problem. For the instance selection, which we mainly concern in this thesis, the algorithm successfully removes the irrelevant features while keeps the informative ones. One main advantage of our approach is that the strangeness has the properties of both filter and wrapper. It is not only simple and can evaluates the relevant of features, but also is a classifier with certain generalization capability. The possible future research direction is to use strangeness to do feature selection in both directions of $F$ simultaneously, which could reduce the complexity of many vision problems dramatically.
Figure 4.5: The most likely selected features for each location.

Figure 4.6: The most likely discarded features by strangeness.
Figure 4.7: The original features detected (left) and the feature selected by strangeness (right).
Chapter 5: Part-based Object Recognition with Strangeness

After the feature selection described in Chapter 4, most local features in training images now have strong relevance with respect to the classification and the complexity of the learning task is highly reduced. In this chapter, we address the problem of modelling and recognizing the objects from the those local features with boosting strategy. The key insight here is to characterize the local classification evidence from each feature with strangeness and integrate it within boosting framework. Note that for face recognition, all extracted local features are relevance and feature selection is not necessary.

5.1 Related Work

Taking a closer look at the extensive body of literature on generic object recognition, many approaches have been proposed to integrate the local evidence and build the final strong classification rules. In this section, we will discuss some of the most relevant and most recent results related to our approach and point out the differences to our method.

Boosting was successfully used by Viola and Jones [87] as the learning ingredient for a fast face and pedestrian detector. The weak classifier was the thresholded average brightness of collections of up to four rectangular regions. The training
images were pre-segmented and the local features were obtained by pixel-by-pixel scanning with different scale and orientation. In our approach, we extract the features by a feature detector. It is able to perform recognition of a wider class of objects on more complicated images. Schneiderman and Kanade [77] also use boosting to improve an already complex classifier and recognize objects from different viewpoints. Kullback-Leibler feature pursuit is used as weak learning rule in the work [52], which need the reliable estimation of several probabilities. Contrary to them, we are using boosting to combine rather simple, non-parametric classifiers. As proved in [23], the generalization capability of the final strong classifier is highly related to the number of weak rules used and the complexity of the weak learner characterized by its VC-dimension. Simpler weak learner can guarantee better generalization power while complex weak learner will lead to over-fitting. Since each local feature is represented in high dimensional space, simple stump function cannot be applied directly. It is also difficult to optimize the KL divergence as shown in [52] due to the small number training images.

Also a wide variety of other learning techniques has been used to solve the task of generic object recognition. Agarwal and Roth [1] use Winnow as the underlying learning algorithm for the recognition of cars from side views. In Winnow, one weak classifier needs to be calculated for each possible threshold a priori. Walraven et al. [88] use Support Vector Machines combined with local features for object recognition. Fergus et al. [18] use a generative probabilistic constellation model for objects with an EM-type learning algorithm. The parts and their constellations can be learned from cluttered images without supervision. We use a model-free approach and propose a variation of standard boosting algorithm, which incorporate the strangeness measure
as the weak learner and learn object parts without supervision. Compared with EM, our learning model has no limitation of the number of parts. It can learn small number of parts for object categories, but also large number of parts for individual objects. Ferrari et al. [19] present an approach where object recognition works even if aggravating factors like background clutter, scale variations or occlusion are very strong. Starting with a small initial set of corresponding features good results are obtained through iterative search. While this work presents a powerful concept of an iterative “active exploration” approach, it is based on a model for an individual object which is learned from pre-segmented training images. Another object recognition approach was introduced by Dorko and Schmid [15]. It is based on the construction and selection of scale-invariant object parts. One classifier is constructed for each object part. The final strong classifier is integrated through the performance of those classifiers on the validation data set using the likelihood ratio and mutual information. It needs clean data for each part to build the classifier, which is often infeasible in real practice.

5.2 Boosting Learning

Our learning model is based on the AdaBoost algorithm [23]. In the training stage labelled images are provided as \((I_1, y_1), \cdots, (I_n, y_n)\) for the learning algorithm, where \(y_i = 1\) if \(I_i\) contains a relevant object and \(y_i = 0\) if \(I_i\) contains no relevant object. The AdaBoost algorithm (see Algorithm 4) was adapted by adding the possibility of putting different weights \(w_i\) on positive and negative training images, and delivers a classifier \(h(I) : I \to y\) to predict whether a given image contains the object or not.

Algorithm 4. Standard AdaBoost Algorithm
Given training images \((I_1, y_1), \cdots, (I_n, y_n)\), where \(y_i = 0, 1\) for negative and positive training images respectively.

- Initialize weight \(w_{1,i} = \frac{1}{m} \frac{1}{l}\) for \(y_i = 0, 1\) respectively, where \(m\) and \(l\) are the number of negatives and positives respectively.

- For \(t = 1, \cdots, T\)
  1. Normalizing the weights, \(w_{t,i} \leftarrow \frac{w_{t,i}}{\sum_{j=1}^{n} w_{t,j}}\), so that \(w_{t,i}\) is a probability distribution.
  2. Training the weak classifier \(h_t(I)\) and evaluate the error with respect to the weights \(w_{t,i}\), \(\epsilon_t = \sum_i w_{t,i} |h_t(I_i) - y_i|\).
  3. If \(\epsilon_t \geq 0.5\), set \(T = t - 1\) and terminate the for loop.
  4. Set \(\gamma_t = \epsilon_t / (1 - \epsilon_t)\).
  5. Update the weight \(w_{t+1,i} = w_{t,i} \gamma_t^{1-e_i}\), where \(e_i = 0\) if image \(I_i\) is classified correctly, otherwise \(e_i = 1\).
  6. Set \(\beta_t = -\log \gamma_t\).

- The final classifier obtained is:
  \[
  h(I) = \begin{cases} 
    1, & \sum_{t=1}^{T} \beta_t h_t(I) \geq \frac{1}{2} \sum_{t=1}^{T} \beta_t \\
    0, & \text{otherwise}
  \end{cases}
  \]

In the AdaBoost algorithm, the weights \(w_{t,i}\) are put on each training image at each round and the weak learner \(h_t\) is constructed which has some discriminative power with respect to the weight \(w_{t,i}\) distribution, that is, at round \(t\)

\[
\sum_{k=1, h_t(I_k) = y_k}^{n} w_{t,k} > \sum_{k=1, h_t(I_k) \neq y_k}^{n} w_{t,k}.
\]  

(5.1)
The best weak learner $h_t$ is not expected to classify the training data well. It should be however better than random chance. In order to boost the weak learner, a sequential learning problem is solved. At each round of learning, the examples are re-weighted according the classification error of the previous weak classifier. The weight $w_k$ is decreased if the prediction for $I_k$ was correct ($h_t(I_k) = y_k$) and increased if the prediction was incorrect. Only the difficult examples, which were incorrectly classified by the previous weak classifier, are emphasized in the current round of learning. The final strong classifier takes the linear combination of weak classifiers followed by a threshold.

Freund and Schapire proved that the training error of the strong classifier approaches zero exponentially in number of rounds [23]. The higher discriminative power the weak learner has, the faster the AdaBoost algorithm converges. In addition, the adaptation of the weights $w_k$ in each round performs some sort of adaptive de-correlation of the weak hypotheses. More importantly a number of results were proved about generalization performance of AdaBoost [76]. With simple and good weak hypotheses, AdaBoost achieves rapidly large margins and low generalization error. For multi-class problem, it can be either reduced to two-class problems using error-correcting codes or treated more directly using multi-class base classifiers. Freund and Schapire [23] proposed two algorithms AdaBoost.M1 and AdaBoost.M2 which are the straightforward generalization of AdaBoost for multi-class base classification. A joint multi-class boosting algorithm was presented in [79] to find the sharing feature among the classes. There are some other multi-class boosting approaches and described in [16].
Obviously AdaBoost is a very general learning technique for obtaining classification functions. To adapt it for a specific application and obtain good generalization performance, suitable and relatively simple weak learner have to be constructed. For the purpose of multi-class visual object recognition, we need use the high dimensional local feature descriptors to construct the multi-class weak classifier. Simple “stump” functions or decision trees on the features cannot be applied directly. As shown in Chapter 2, the strangeness is the non-parametric measurement capturing the discriminative ability of the features and is applicable for both two-class and multi-class problems. The higher the strangeness, the higher is the uncertainty associated with the label and the usefulness of the feature for accurate discrimination decreases. In next section, we will show how the strangeness is used to construct the weak learner in standard boosting procedure.

5.3 Non-parametric weak learner

After feature selection, each training $I_k$ image is represented by the selected feature set $\{g^k_j\}$. The strangeness of each feature is computed according to Equation 2.9. Considering strangeness as the base classifier, we can apply the AdaBoost algorithm on the selected feature set directly. However, the feature selection algorithms proposed in Chapter 4 only discard irrelevant features and the redundant features survive in the database. As a result, several features may be extracted from the almost same location of the same object thus the redundancy information exists. If each feature is considered as a weak classifier as in [69], the final strong classifier will be over-fitting and have the low generalization capability. For example, eye is a very important feature to distinguish face from other objects. If the final strong classifier has several
“eye” weak learners associated with large coefficient $\beta_i$, it has high probability of misclassifying the test face if the “eye” feature is not detected in the image.

![Figure 5.1: The object parts grouped – the weak rule in Boosting.](image)

To reduce the redundant information among features and achieve high generalization ability of the final classifier, we apply $K$-means clustering algorithm on the selected feature set of each class to group them into $P$ parts. Each part is now considered as a weak classification rule which is used as the base classifier in AdaBoost. Fig. 5.1 shows the parts of the face and motorcycle categories after clustering. For face recognition, the local features are extracted in different scales and bandwidth (see Section 3.3). To reduce the redundant information we group the features with the same location and scale as one part and consider it as the weak classification rule in the AdaBoost. Thus the face is represented by $43N_s$ parts, each of which has $N_b$ feature instances $\{g_j^i\}_{j=1}^{N_b}$ extracted in different bandwidth channel. There are two kinds of benefits for such representation. At first, it reduces the redundant information among features thus there is little redundant information among weak hypotheses, which improves the generalization performance of the final strong classifier. Second, it constrains the maximal number of the weak rules in the final linear combination, thus the total computational complexity of the learning algorithm. Based on recent
research on boosting learning [57], AdaBoost will be over-fitting if the number of weak learners is too large. Constraining the part number $P$ in certain range will reduce the chance of over-fitting. Since most features are now relevant, the clustering algorithm and the number of clusters do not have much influence on the performance of the location and object category recognition. As shown in Section 5.5, the performance of final classifier has close performance when $P$ is in some range.

From the training data set, each object $c$ is consequently represented by $P$ parts, each of which has $N_i$ feature instances $G^c_i = \{g^i_j\}_{j=1}^{N_i}$. Instead of parametric modelling of these parts and building the classification boundary, as did in many previous works, we keep them and the corresponding feature instances as training gallery. We then apply the base classifier on $P$ parts and learn the coefficients and thresholds of weak learners through a validation data set. Given the validation image $V^i$ and its local features descriptor $\{g(V^i)_j\}$ with putative object label $c$, the matched features $\{\hat{g}(V^i)_j\}_{j=1}^{P}$ are found which are the closest feature from $\{g(V^i)_j\}$ to each part of class $c$ in the gallery. Then the strangeness $\{\alpha^c_i\}$ of $\{\hat{g}(V^i)_j\}_{j=1}^{P}$ are computed with the assumption of putative class $c$. With $C$ classes in the training gallery, $C$ group of strangeness are obtained for each validation image. If $M$ validation images are given for each class, for each part of each class, we have $M$ positive strangeness measures and $M(C - 1)$ negative ones. Our weak hypothesis is to select the matched feature $\{\hat{g}(V^i)_j\}_{j=1}^{P}$ and the strangeness threshold $T_j$ for each part of the class. The algorithm of weak learner is shown in Algorithm 5.

**Algorithm 5. Strangeness Weak Learner**
• **Input**: Training gallery \( \{ G^c_j \}_{j=1}^{P}, c = 1, \cdots, C \), where \( G^c_j \) is the feature instance set of \( j \)th part of class \( c \), and validation images \( \{ V_i, i = 1, \cdots, MC \} \) and associated feature \( \{ g(V_i) \} \).

• **Distance metric function**: Let \( d(., .) \) be the distance between two feature descriptors.

• **Strangeness computation**: For each part \( j \) of class \( c \), find the nearest feature \( \tilde{g}(V_i)_j \) between \( \{ g(V_i) \} \) and \( G^c_j \). The strangeness of \( \tilde{g}(V_i)_j \) is then computed based on Equation 2.9 with the putative the class label \( c \) of \( V_i \). Each part of the class \( c \) now has \( MC \) strangeness \( \{ \alpha^c_{k,j} \}_{k=1}^{MC} \), \( M \) of which are positive and \( M(C - 1) \) are negative.

• **Strangeness sorting**: For each part \( j \) of class \( c \), let \( \pi(1), \cdots, \pi(MC) \) be the permutation such that:

\[
\alpha^c_{\pi(1)} \leq \alpha^c_{\pi(2)} \leq \cdots \leq \alpha^c_{\pi(MC)}.
\]

• **Select the threshold of weak learner**: For each part \( j \) of class \( c \), find the best index position \( s \) such that the maximal classification rate is achieved:

\[
rate(j) = \max_s \sum_{k=1}^{s} w_{\pi(k)} 1(\alpha_{\pi(k)})
\]

where \( 1(\alpha_{\pi(k)}) \) is 1 if \( \alpha_{\pi(k)} \) is positive and 0 otherwise. Then the threshold of current weak learner is:

\[
\theta(j) = \frac{\alpha_{\pi(s)} + \alpha_{\pi(s+1)}}{2}.
\]
• **Select the best weak learner**: Find the best part \( m = \max_j \text{rate}(j) \). Then the best weak learner of current round is the \( m \)th part with the best threshold \( T_m = \theta(m) \). Update the weight \( w_k \) and compute the coefficient \( \beta_t \) according to error \( 1 - \text{rate}(m) \).

The Strangeness Weak Learner is model-free and non-parametric and as simple as the stump function. The main computational burden is the calculation of strangeness of \( g(V_i)_j \) with putative label \( c \), since it needs the distances from \( g(V_i)_j \) to all features in the training gallery. However, such computation can be done prior to boosting and weak learner finder. Given the parameter \( k \) for strangeness computation, \( k \) nearest neighbor distances from \( g(V_i)_j \) to each part of each class are computed first and stored. The strangeness of \( g(V_i)_j \) with any putative label can be calculated from those distances. After all strangeness are computed, the remaining calculations in boosting learning are very inexpensive.

## 5.4 Learning Model and Recognition

With the weak learner in Algorithm 5, the standard AdaBoost algorithm (see Algorithm 4) is applied directly. The learning diagram is shown in Fig 5.2. In this learning algorithm, each round of boosting is close to the hypothesis test, similar to transductive inference [48, 75, 83]. In each iteration, the hypothesis \( H_c \) is made first (putative label \( c \) for the image) and the strangeness is calculated under such hypothesis. The final output of boosting is used to test if such hypothesis is valid or not. Drawing an analogy between weak classifiers and features, this learning algorithm is another aggressive feature selection mechanism for selecting a small set of “good” features which nevertheless have significant variety. Although each class is represented by \( M \)
parts, only some of them are useful after boosting learning. Some parts may have no discrimination power with respect to the weight \( w_t \) at round \( t \) and have the zero coefficients in the final linear combination. Finally \( C \) group of coefficients \( \{\beta_i^c\}_{t=1}^P \) are obtained, which tell the importance of each part for each object. The coefficients are then normalized such that \( \sum_{i=1}^P \beta_i^c = 1 \) for each object.

Figure 5.2: Boosting Learning with strangeness measure as the weak classifier.

Recognizing query image \( Q \) is straightforward, similar to the validation stage, as shown in Fig. 5.3. With putative label \( c \) for \( Q \), the strangeness are computed, compared with the thresholds \( T_i^c \) and the posterior \( P(c|Q) \) is estimated through the coefficients \( \beta_i^c \). The label of \( Q \) is predicted by \( \arg \max P(c|Q) \). Compared with most parametric approaches, the computational complexity of the testing stage is relatively expensive since we need compute the distance between the feature \( g(Q)_j \) and all features in the training gallery. For this step however more efficient approximate distance computation schemes can be employed using \( kd \)-tree. However, our learning approach is model-free, without complex parameter estimation and search for correct model in large space of models, and works for any number of parts in object representation.

There are two parameters in our learning approach which need to be tuned: the
number of nearest neighbor $k$ in strangeness measure and the number of parts $P$ for weakly supervised learning. $k$ can be estimated from the number of training images for each class and the repeatability of the feature detector. We can vary $k$ and find the best one with best performance using the validation data set. We also can vary $P$ and find the best one through validation. As demonstrated in Section 5.5, the performance of final classifier is very stable and has close performance when $P$ is in some range.

### 5.4.1 Weakly Supervised Learning

Recent object recognition approaches, which used weakly supervised training data (e.g. [15,17,69]), achieved good recognition performance while disambiguating object categories from background. The training images are given with positive objects as well as the cluttered background. Only the label of the whole image is given while the location of the object in the image is unknown. We need to learn the representation of objects and recognize the objects directly without prior segmentation. We instead of just discriminating the object category from the background images in [15,17,18,69]
propose a two-stage hierarchical boosting learning, which can distinguish the object from both the background and other objects, similar to the open set learning in [48]. At first, selected features are applied between objects and background a two-class boosting learner is learned to classify each object category from background images. This is the object detection stage. Based on the features selected in the first stage, further feature selection is done and another one-vs-all boosting learner is used to classify different object categories. This is the close set multi-class classification stage. The diagram of hierarchical approach is shown in Fig. 5.4. In the second stage, the label of a query image $Q$ is predicted by $\arg\max P(c|Q)$. It is necessary to use these two stages. Since the background features are uniformly distributed in the feature descriptor space, we cannot model the background with parts. As a result we cannot reliably estimate $P(\text{background}|Q)$. Given the estimated $P(c|Q)$, it is very hard to find a threshold $\tau$ such that $Q$ is background if $\max P(c|Q) \leq \tau$. To avoid estimating such threshold, the boosting is proposed to deal with the background. The high performance can be achieved in this stage. Some of them have no discriminative power for classifying the object categories but have the ability to distinguish the objects from background.
5.5 Experimental Setup and Results

In this section we present numerical evaluation of our described method in the application areas of both supervised object recognition and weakly supervised object category recognition. For supervised object recognition, we apply it on location recognition for robot navigation task, where the background clutter is not so prominent, and face recognition. For each task, the performance is evaluated and compared with other approaches.

5.5.1 Location Recognition

We first test our approach on location recognition task. In this task, the video sequence is partitioned into 18 different locations, each of which is represented by 4-7 number of views. Each individual view is represented by SIFT keypoints. We have three video sequences. The representative views for each location are sampled from the first sequence. The remaining frames of first sequence are used as the validation data set. The other two sequences are used for testing. More detailed information about the data and representation can be found in [46]. Fig 5.5 shows the representative view for some locations, which demonstrate the variability of the data set.

<table>
<thead>
<tr>
<th>sequence (# of frames)</th>
<th>voting on original set [46]</th>
<th>boosting approach</th>
</tr>
</thead>
<tbody>
<tr>
<td>No.1(296)</td>
<td>100.0%</td>
<td>100.0%</td>
</tr>
<tr>
<td>No.2(134)</td>
<td>82.1%</td>
<td>86.8%</td>
</tr>
<tr>
<td>No.3(130)</td>
<td>83.1%</td>
<td>87.8%</td>
</tr>
</tbody>
</table>
On average there are about 450 features detected for each frame. After feature selection using strangeness, there are about 10% discriminative features selected. In our experiments, $\gamma = 0.9$ and $k = 3$ in feature selection Algorithm 3. Because each feature has different scale, the selected features characterize the local and global discriminative information of that location. Table 5.1 shows the recognition rates of our approach with $P = 20$ compared with the voting approach using the original detected features (see [46]). The presented approach performs better, but at same time reduces the matching computational cost by about 90%. Very close performance is achieved if number of parts $P$ for each location varies from 15 to 30.

To deal with the dynamic changes in the environment, we further incorporate additional knowledge about neighborhood relationships between individual locations in a Hidden Markov Model (HMM), as did in [41]. The rationale behind this choice is, that despite the presence of ambiguities in recognition of individual views the temporal context should be instrumental in resolving them. The output of final boosting classifier is approximated as the observation likelihood and about 97% recognition
Figure 5.6: Classification results with for Test Sequence 1 and Sequence 2 considering the spatial relationships modelled by HMM. The black circles correspond to the location labels assigned to individual frames of the video sequence.

rates are achieved for both test sequences. Figure 5.6 shows the location label which has the highest probability in the model.

5.5.2 Object Category Recognition

We further tested our approach for weakly supervised object category recognition on the database used by Fergus et al. [18]. The images in the database contain objects with arbitrary scale and location as well as the highly textured background clutter. In order to compare the performance, we use 4 categories: motorbike, airplane, faces and cars(side). The database also has the background images, which are used to learn the object models and test the performance. Different from the experiments in [18] and [69], for each object we randomly sample 30 images as the training gallery, 30 images as validation data set and use the remaining images for testing as well as the background images. The features are detected by affine covariant regions and represented by SIFT descriptor. The hierarchical boosting learning described in
Section 5.4.1 was applied on the database.

The first stage of hierarchical learning discriminates the objects from background as did in [18, 69]. Fig. 5.7(a) shows the ROC curves of our approach on the database where $\gamma = 0.9$, $P = 30$ and $k = 5$. Fig. 5.7(b) shows the performance with respect to the number of clusters $P$. Table 5.2 shows the equal error rates of our approach compared with the other two approaches. Our method is a little better than both of them except for the faces. From the results in Fig. 5.7(b) we can see our approach is very stable when the number of clusters $P$ is in the range $[25, 50]$. When $P$ is small, too little evidence is integrated from the local parts and the final strong classifier has no enough discriminant power. When $P$ is too large, similar features may have multiple clusters and redundant information exists between weak hypotheses; thus the final strong classifier will be over-fitting. Fig. 5.8 and Fig. 5.9 show the examples of best feature, that is, the part with largest coefficient after boosting learning, for face and motorbike to discriminate from background in the first stage. It clearly demonstrates that our approach successfully extracts the discriminative and meaningful features from images without prior segmentation.

Table 5.2: The ROC equal error rates on the database used by Fergus et al. [18]

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Our approach</th>
<th>Fergus et al. [18]</th>
<th>Opelt et al. [69]</th>
</tr>
</thead>
<tbody>
<tr>
<td>Motorbikes</td>
<td>96.1%</td>
<td>92.5%</td>
<td>92.2%</td>
</tr>
<tr>
<td>Faces</td>
<td>94.4%</td>
<td>96.4%</td>
<td>93.5%</td>
</tr>
<tr>
<td>Airplanes</td>
<td>93.7%</td>
<td>90.2%</td>
<td>88.9%</td>
</tr>
<tr>
<td>Cars(side)</td>
<td>93.1%</td>
<td>88.5%</td>
<td>83.0%</td>
</tr>
</tbody>
</table>

In the second stage, a one-vs-all boosting classifier is learned on the features selected in the first stage. It distinguishes each object from all other object categories,
Figure 5.7: (a) The ROC curve for image classification on the faces, motorbikes, airplanes and cars(side) data set used by Fergus et al. [18]. (b) The equal error rates with respect to the number of clusters $P$.

not just at the level of chance as showed in Table 2 in [18]. The work in [15,17,69] did not report how their approaches perform on separating one category from the others. Table 5.3 presents the performance of our learning approach across the four classes. Very good recognition rates are achieved. The model for each object successfully rejects the input images from other objects. In this thesis, the initial features in this stage are the features selected in the first stage. The performance can be improved if we locate the objects first through the results in the first stage and use all features on the objects to learning the multi-class classifier.

In addition to those four categories, we add another three categories, cars(rear), cougar faces and watches from Caltech 101 category data base [17]. Table 5.4 presents the recognition rate in two stages. Very good performance is achieved. If we take a look at the misclassified examples, most of them have very low contrast between objects and background thus few features are detected on the objects. Some examples
Table 5.3: The performance of the final strong classifier at the second stage on the database used in [18].

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Motorbikes model</th>
<th>Faces model</th>
<th>Airplanes model</th>
<th>Cars(side)model</th>
</tr>
</thead>
<tbody>
<tr>
<td>Motorbikes</td>
<td>93.1%</td>
<td>2.5%</td>
<td>1.6%</td>
<td>2.8%</td>
</tr>
<tr>
<td>Faces</td>
<td>1.1%</td>
<td>93.4%</td>
<td>4.5%</td>
<td>1.0%</td>
</tr>
<tr>
<td>Airplanes</td>
<td>2.0%</td>
<td>0.0%</td>
<td>95.4%</td>
<td>2.6%</td>
</tr>
<tr>
<td>Cars(side)</td>
<td>2.1%</td>
<td>0.0%</td>
<td>6.9%</td>
<td>91.0%</td>
</tr>
</tbody>
</table>

are shown in Fig 5.10. More robust and reliable feature detectors are needed to overcome this phenomenon.

Table 5.4: The performance of cars(rear), cougar faces and watches.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Cars(rear)</th>
<th>Cougar faces</th>
<th>Watches</th>
</tr>
</thead>
<tbody>
<tr>
<td>First Stage</td>
<td>88.3%</td>
<td>95.1%</td>
<td>90.2%</td>
</tr>
<tr>
<td>Second Stage</td>
<td>90.3%</td>
<td>94.4%</td>
<td>92.1%</td>
</tr>
</tbody>
</table>

### 5.5.3 Face Recognition

We carried out our experiments using frontal face data from the University of Notre Dame collected during 2002 – 2003 academic year and now part of the Face Recognition Grand Challenge (FRGC) face image database [72]. The face images were acquired under different lighting conditions and with different facial expressions. In addition there is temporal variations as the face images were acquired at different sessions over one year period. At each session, some subjects have two more images which were taken in the hallway outside the studio where the background is cluttered and lighting is uncontrolled. We sampled 200 subjects from the whole data set, each
Figure 5.8: Examples of the best feature for face after boosting learning.

of which has 48 images in total and 16 ones among them taken in uncontrolled environment. Each image is normalized with size $289 \times 289$ and 256 gray scale levels, as shown in Fig. 5.11.

For each face images, the local patches are extracted and the corresponding SIFT descriptors are computed as described in Section 3.3. In our experiments, $N_s = 5$ and $N_b = 4$. Therefore each face is represented by $P = 43N_s = 215$ parts, each of which $N_b + 1 = 5$ feature instances. If we consider the symmetry of the frontal face, there are $P = 26N_s = 130$ parts for each face, each of which has different number of feature instances. For each subject, we randomly select 12 images as training set, another 12 images as the validation set and the remaining 24 images as testing set. Euclidian distance is used to compute the strangeness with $k = 5$.

In order to evaluate the performance of our approach, we construct a voting recognition approach-Transductive Confidence Machine with strangeness [42,48], similar to
the framework in [34]. Transductive Confidence Machine approximates the p-Value using the strangeness and predicts the class membership with the largest p-Value. More detail information are shown in Chapter 6. Let the feature instances of $j$th part of class $c$ be $\{g_{jl}^c\}_{l=1}^{M_c}$, where $M_c$ is the number of features in part $j$. The corresponding strangeness $\{\alpha_{jl}^c\}$ are computed based on the Equation 2.9. Given the test faces

Figure 5.9: Examples of the best feature motorbike after boosting learning.

Figure 5.10: Examples of misclassified image where few features are detected on the objects.
image $T_i$ with its local features $\{g(T_i)_j\}$ of part $j$, similar to the strangeness weak classifier (see Algorithm 5), the matched features $\{\tilde{g}(V_i)_j^c\}_{j=1}^{P}$ are found with putative face label $c$. The strangeness $\{\alpha(T_i)_j^c\}$ of those matched feature in $T_i$ are then computed under such assumption. Note that $\alpha_{jl}^c$ need recompute if the identity of their $k$ nearest neighbor features changes due to the location of $\tilde{g}(V_i)_j^c$. One chooses that particular labelling from $C$ classes driven by the largest p-value for class membership.

After the label of each part of test face $T_i$ is obtained, the class membership of $T_i$ is predicted by the class voting, i.e., the putative label $c$ has the largest number of matched parts. Validation data is not needed in this voting approach.

Table 5.5 shows the top-1 rank identification rates using strangeness-based boosting learning model and voting approach, respectively. Our approach performs quite well and much better than the alternative voting approach. The introduction of 2nd order patches increases somehow the performance.

Fig. 5.12 shows the average coefficients of 1st order parts in the final Boosting classifier, which yields the importance of local region to distinguish the subject from others. In the figure, the parts are scanned from left to right and from top to bottom as shown in Fig. 3.5(a). From the figure we can see that averagely the regions around
Table 5.5: Top-1 rank performance with different learning approaches.

<table>
<thead>
<tr>
<th>Representation</th>
<th>Without symmetry</th>
<th>Consider symmetry</th>
</tr>
</thead>
<tbody>
<tr>
<td>Voting</td>
<td>1st order patches only</td>
<td>1st and 2nd order patches</td>
</tr>
<tr>
<td>Approach</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Strangeness-based</td>
<td>87.8%</td>
<td>90.3%</td>
</tr>
<tr>
<td>Boosting</td>
<td>88.1%</td>
<td>89.2%</td>
</tr>
</tbody>
</table>

The nose contribute more than other regions to discriminate the faces. For each location of the face, the median size regions is more important than the small and large size regions.

![Coefficients distribution considering symmetry](image1)
![Coefficients distribution without symmetry](image2)

(a) Considering symmetry  (b) Without symmetry

Figure 5.12: The average coefficients of the 1st order parts in final Boosting classifier.

In order to evaluate the performance of our approach with respect to the occlusion, we modify the test image to simulate occlusion in the following way. A circle region with radius $r$ is randomly chosen in the face image, the content of which is either set 0 or filled with random pixel values in $[0, 255]$. Fig. 5.13 shows the samples of occluded images. The average top-1 rank recognition rates with respect to $r$ are
shown in Fig. 5.14 where the occluded regions are either totally blank or randomly filled with pixel values. From the results, we can see considering symmetry can increase the performance. On the average the recognition rate decreases when the radius of occluded region increases but it does not drop too much. The performance are very stable when the occluded regions are not too large. If voting approach in is applied, the recognition rate is a little bit lower than strangeness-based boosting learning and the performance drops dramatically with respect to $r$.

![Figure 5.13: The sample occluded images.](image1)

(a) $r = 20$  
(b) $r = 40$

Figure 5.14: The sample occluded images.

![Figure 5.14: Average top-1 rank recognition rates with respect to $r$.](image2)

(a) Occluded regions are filled by 0.  (b) Occluded regions are randomly filled.

The occluded regions are randomly chosen in the above experiment. The next
experiment considers the case when the occluded regions are fixed, e.g., eyes, nose and mouth. Fig. 5.15 shows the recognition rates with respect to radius $r$ when one eye, nose or mouth are occluded and symmetry is considered. The performance are almost same when one eye is occluded and the other eye is available. From the figure we can see that the occlusion of nose affects the performance more than the mouth and eyes. It is consistent with the observation from the distribution of parts’ coefficients in Fig. 5.12.

Figure 5.15: Recognition rates with respect to $r$ when eye, nose or mouth is occluded.

(a) Occluded regions are filled by 0. (b) Occluded regions are randomly filled.

5.6 Discussion

In this chapter we propose a novel object recognition approach, assuming that objects are represented by several local features (parts), whose descriptors are invariant with respect to certain variation in object appearance. It works for individual objects and object categories recognition. After a simple and efficient feature selection algorithm is applied if necessary, a new non-parametric weak leaner is employed in the boosting framework. Specifically the strangeness measure is used to quantify the
uncertainty of features as the base classifier. The approach is applied and validated on supervised location recognition, face recognition and weakly supervised object category recognition problems. It successfully characterizes the locations locally and globally. For the face recognition, the approach is shown to be robust with respect to facial, illumination, and temporal changes, on one side, and missing local components, on the other side. For the last task, we propose a two-stage hierarchical approach and learn how to distinguish the object from both background and other objects. Experiments are carried out using Caltech object category database provided by [17]. While testing the simpler presence and absence on an object, our approach yields results comparable to the state-of-the-art. The performance on a considerably harder multi-class classification task, demonstrates the effectiveness of the proposed method. The presented flexible approach is applicable for general classification problems and constitutes a significant step towards generic object recognition and local evidence integration.

There are several advantages over our approach: no complex parametric model assumption and parameter estimation; the limitation of the feature detector’s repeatability compensated by boosting learning stage and applicable for both two-class and multi-class classification. There are of course some limitations. The assumption that the objects are represented by parts may be invalid, for example, the wired objects, bottles, trees. For large number of classes and training image, efficient algorithm is needed to compute k-nearest neighbors. When the local features are obtained by some detectors, the performance of the algorithm depends on the repeatability of the feature detectors. Our approach could be improved when multiple types of features are used.
Chapter 6: Open Set Recognition using Strangeness

6.1 Background

The generic open set recognition problem, a major challenge in pattern recognition, operates under the assumption that not all the test examples have mates (counterparts) in the training data set of known subjects, as illustrated in Fig. 6.1. It requires the a priori availability of a reject option to answer “none of the above”. If the probe is detected rather than rejected, the recognition engine must then identify / recognize the subject. The operational analogue for open set recognition is the (usually small) Watch List, Intrusion detection or Surveillance task, which involves (i) negative identification (“rejection”) due to the obvious fact that the large majority [almost all] of the subjects screened, e.g. people and internet connection, are law abiding, and (ii) correct identification for those that make up the watch list or intrusion.

![Diagram of open set and watch list tasks](Figure 6.1: Generic Open Set and Watch List Tasks.)
The open set recognition engine, however, require significant tuning and calibration, for setting the detection thresholds among other things, before “lug and play”. Setting thresholds is not easy to automate due to their strong dependency on data quality and the composition of training data. Note also that “much more is known about the population, [or genuine customers,] of an application than is known about the enemies, [i.e., the imposters that have to be rejected]. Consequently, the probability of a false alarm rate (FAR), a false match [for screening and positive identification], is hard to estimate. Hence, the false reject rate (FRR), that concerns open set negative identification, is easier to estimate than the false alarm rate, because the examples of the enemy population are not available” [13]. The thresholds needed for field deployment have to be set up ahead of time, i.e., a priori, and without resorting to additional client, e.g., impostor data. The alternative of setting the thresholds a posteriori using the ground truth available from the aggregate similarity scores recorded for matching the probe set against the gallery set is not appropriate because the ground truth is not available in real practice. In this chapter, we propose a method to learn the threshold a prior using strangeness and demonstrate its performance in the face recognition system, in particular.

6.2 Open Set Face Recognition

Conventional threshold selection methods typically compute the distribution of inter- and intra subject distances, and then choose a threshold to equalize the overlapping areas of the distributions, i.e., to equalize the false acceptance rate (FAR) and false rejection rate (FRR), as did in [26, 29, 73]. “The success of this approach, however, relies on whether the estimated distributions match the subject- and impostor-class
distributions. Session-to-session subject variability, however, contributes much bias to the thresholds, rendering the authentication system unstable" [10].

As discussed in Chapter 2, the strangeness measure function defines a space where Martin-Löf randomness test [49] can be applied (see Equation 2.3). The values taken by such randomness tests are called p-values. Given \( l \) training examples associated with strangeness \( \alpha_i \) and the strangeness \( \alpha_{new}^c \) of new test exemplar \( e \) with putative label \( c \), a valid randomness test approximation [59] for this transductive (putative label \( c \)) hypothesis is constructed as:

\[
p_c(e) = \frac{\#\{i: \alpha_i \geq \alpha_{new}^c\}}{l + 1}
\]  

(6.1)

An alternative valid randomness approximation [75] and the one that we use here defines the p-value for the test exemplar \( e \) (with putative label \( c \)) as:

\[
p_c(e) = \frac{f(\alpha_1) + f(\alpha_2) + \cdots + f(\alpha_l) + f(\alpha_{new}^c)}{(l + 1)f(\alpha_{new}^c)}
\]  

(6.2)

where the function \( f \) is monotonic non-decreasing with \( f(0) = 0 \). Experimental data reported later in this chapter uses \( f(\alpha) = \alpha \). Our empirical evidence has shown that the alternative randomness approximation 6.2 yields better performance than the standard one 6.1, which may suffer from "distortion phenomenon" [75]. If there are \( C \) classes in the training data, there are \( C \) p-values for each test exemplar \( e \). Using p-values one chooses that particular labeling driven by the largest randomness deficiency for class membership, i.e., the putative label \( y \) that yields the least strangeness or correspondingly the largest p-value. This largest p-value is also defined as the credibility of the label chosen, which is a measure of information quality. The associated
confidence measure, which is derived as the 1st largest p-value (or one) minus the 2nd largest p-value, indicates how close the first two assignments are. The confidence value indicates how improbable classifications other than the predicted labeling are, while the credibility value shows how suitable the training set is for the classification of that working exemplar. Note that each new test exemplar \( e \) with putative label \( c \) requires to recompute, if necessary, the strangeness for all the training exemplars when the identity of their \( k \) nearest neighbors exemplars changes due to the location of (the just inserted new exemplar) \( e \).

Based on the p-values defined above, Proedrou et al. [42] have proposed the TCM-kNN (Transduction Confidence Machine - k Nearest Neighbor) to serve as a formal transduction inference algorithm for classification purposes. TCM-kNN does not address, however, the detection (decision) aspect needed for open set face recognition. Our proposed solution for the detection aspect involves using the \( PSR \) (peak-side-ratio) that characterizes the distribution of p-values. It implements the equivalent of the likelihood ratio used in detection theory and hypothesis testing, where likelihood ratio is the ratio between the hypothesis \( H_0 \) that the unknown probe \( e \) belongs to the gallery and \( H_1 \) (alternative hypothesis) that it does not belong.

The distribution for the \( PSR \), if impostor cases were made available, serves to determine how to threshold in order to accept or reject a particular test exemplar \( e \). Towards that end, one would relabel the training exemplars, one at a time, with all putative labels except the one originally assigned to it. The corresponding \( PSR \) should resolve each such relabeled exemplar suitable for rejection because its new label is mistaken. The resulting distribution for the \( PSR \) determines then when to reject working exemplars as impostors. Open Set TCM - kNN implements the
above concepts and it is described as following.

6.2.1 Open Set TCM-kNN Algorithm

Open Set recognition operates under the assumption that not all the probes have mates in the gallery and it thus requires the reject option. Given a new test exemplar, the p-values output from Open Set TCM-kNN records the likelihoods that the new exemplar comes from each putative subject in the training data. If some p-value is high enough and it significantly outscores the others, the new exemplar can be mated to the corresponding subject with credibility p. If the top ranked (highest p-values) choices are very close to each other and outscore the other choices, the top choice can still be accepted but its recognition is questionable due to ambiguity and yields low confidence. The confidence measures the difference between the 1st and 2nd largest (or consecutive) p-values. If all p-values are randomly distributed and no p-values outscore other p-values enough, any recognition choice will be questionable and the new exemplar should be rejected. The proposed PSR (peak-to-side ratio)

\[
PSR = \frac{p_{\max} - p_{\text{mean}}}{p_{\text{stdev}}} \quad (6.3)
\]

characterizes those characteristics of p-value distribution, where \(p_{\text{mean}}\) and \(p_{\text{stdev}}\) are the mean and standard deviation of the p-value distribution without \(p_{\max}\).

The threshold for rejection is learned a priori from the composition and structure of the training data set at enrollment time. Each training exemplar \(t\) is iteratively reassigned to all possible classes but different from its own and the p-values are recomputed accordingly. The PSR is derived using the recomputed p-values with \(t\) playing the role of an impostor. The PSR values found for such impostors are low
(since they do not mate) compared to those derived before for legitimate subjects and they require rejection. The PSR distribution provides a robust method for deriving a priori the operational threshold $\Theta$ for detection as

$$\Theta = PSR_{mean} + 3 \times PSR_{stdev}$$

(6.4)

where $PSR_{mean}$ and $PSR_{stdev}$ (standard deviation) are characteristic for the PSR distribution. The probe is then rejected if the relationship $PSR_{new} \leq \Theta$ holds true. Correspondingly, authentication takes place for (large) values exceeding $\Theta$.

There are conceptual similarities between the use of the PSR to approximate the likelihood ratio and scoring normalization methods used in speaker verification [26, 73]. The alternative hypothesis for speech is modelled using either the cohort or the universal background model (UBM). The cohort approximates the alternative $H_1$ hypothesis using speech-specific (same gender impostor) subjects, while UBM models $H_1$ by pooling speech from several speakers and training a single speaker background model. The PSR measure is conceptually related to the cohort model, as both implement likelihood ratio using local estimation for the alternative hypothesis. The ability of the cohort model to discriminate the speaker’s speech from those of similar, same gender impostors is much better than that offered by UBM [55] and it leads to improved security at lower FAR (false acceptance rates). Similar arguments hold for face recognition.

In order to evaluate and compare the performance of Open Set TCM-kNN algorithm, Open Set standard classifier is presented, which derive the rejection threshold from the intra- and inter-distance (similarity) distribution of training exemplars in a fashion similar to that used by FRVT2002 [70]. The statistics of intra-distance
(“within”) distribution set the lower bound of the threshold and the statistics of inter-distance (“between”) distribution set the upper bound. As the minimum distance of the new (test / probe) exemplar to the prototypes for each class becomes closer to or larger than the upper bound, the more likely the new testing exemplar will be rejected. Our experiments have shown that face recognition performance varies according to the threshold chosen.

6.3 Experimental Results

The data set (see Fig 6.2) from FERET [71] consists of 750 frontal face images corresponding to 250 subjects. 200 subjects come from the difficult batch #15 that was acquired using variable illumination and/or facial expressions, while the remaining different 50 subjects consists of are drawn from other batches. Each subject has three normalized (zero mean and unit variance) images of size 150 x 150 with 256 gray scale levels. Each column corresponds to one subject. The normalized 300 face images from 100 subjects are used to generate PCA and FLD (Fisher Linear Discriminant) face basis. 50 subjects are randomly selected from batch #15 and the remaining different 50 subjects were drawn from other batches. The remaining 450 face images for 150 subjects are used for enrollment and testing. They are projected on the PCA and FLD face bases derived ahead of time to yield 300 PCA coefficients and 100 Fisher-faces using FLD on the reduced 300 PCA space [51]. For each subject, two images are randomly selected as training and the third one as testing.

We have used several well-known similarity measures to evaluate their effect on different face representation (PCA and Fisherfaces) when using TCM-kNN (k=1). Our empirical findings indicate that Mahalanobis + $L_2$ distance is superior to others
when expressive features (driven by PCA) are used, while cosine is superior when discriminating (Fisherfaces) features are used. Only those two distance measures are used in the later experiments. Given two \(n\)-dimensional vectors \(X, Y \in \mathbb{R}^n\), the distance measures used are defined as follows:

\[
\begin{align*}
    d_{\text{Mah}+L_2}(X, Y) &= (X - Y)^T \Sigma^{-1}(X - Y) \\
    d_{\text{cos}}(X, Y) &= -\frac{X^T Y}{\|X\|\|Y\|}
\end{align*}
\]

(6.5) (6.6)

where \(\Sigma\) is the covariance matrix of the training data. For PCA, \(\Sigma\) is diagonal and the diagonal elements are the (eigenvalues) variances of the corresponding components.

### 6.3.1 Open Set Recognition Results

The experimental results of the algorithms in Section 6.2 are reported in this section. The recognition rate is the percentage of subjects whose probe is correctly recognized or rejected. Faces are represented using either 300 PCA or 100 Fisherfaces components. From the 150 subjects available, 80 subjects are randomly selected to form a
fixed gallery, while another 80 subjects are randomly selected as probes such that 40 of them have mates in the gallery, i.e., the gallery and probe sets have an overlap of 40 subjects. The gallery consists of two (out of 3) randomly selected images; while the probes consist of the remaining one (out of 3) images for faces that belong to the gallery and one (out of 3) randomly selected image for subjects that do not belong to the gallery. The same experiment is run 100 times for different probe sets. Fig. 6.3 shows the mean recognition rate for different thresholds. When ground truth is available the thresholds Θ are optimally set to yield maximum performance, and the reject decision is taken if \((\min d) > Θ_{\text{reject}}\). The best average (over 100 experiments) authentication (correct rejection and identification) rates (see Fig. 6.3) for Open Set PCA, Fisherfaces classifiers that yield \(\text{FAR} = 7\%\) are:

- \(74.3\% (\text{s.d.}=3.06\%)\) for PCA representation and sometime the optimal \(Θ \sim (\text{Intra}_{\text{mean}} \times \text{Intra}_{\text{stdev}} + \text{Inter}_{\text{mean}} \times \text{Inter}_{\text{stdev}})/(\text{Intra}_{\text{stdev}} + \text{Inter}_{\text{stdev}})\),

- \(85.4\% (\text{s.d.}=2.30\%)\) for Fisherfaces representation and sometime the optimal \(Θ \sim (\text{Intra}_{\text{mean}} \times \text{Inter}_{\text{stdev}} + \text{Inter}_{\text{mean}} \times \text{Intra}_{\text{stdev}})/(\text{Intra}_{\text{stdev}} + \text{Inter}_{\text{stdev}})\).

For Open Set PCA, the results are very close if the number of components used varies from 150 to 300, while for Open Set Fisherfaces, the results are very close if the number of components used varies from 55 to 90. More experiments have been done randomly varying the gallery set and similar results were obtained. The optimal threshold, however, varies largely with the gallery set and probe, and would be hard to be determined a priori. Attempts made to learn the threshold a priori, i.e., without ground truth knowledge were unsuccessful.
The same Open Set experiment was run then using Open Set TCM-kNN for \( k = 1 \). The only difference now is that the rejection threshold \( \Theta \) is computed a priori according to the PSR procedure described in Section 6.2.1 (see Fig. 6.4) Authentication is driven by large PSR and the average authentication (correct rejection and identification) rates for (measured) FAR = 6% are:

- 81.2%(s.d.=3.1%) for PCA using \( \Theta = 5.51 \) and the \( Mah + L_2 \) distance,
- 88.5%(s.d.=2.6%) for Fisherfaces using \( \Theta = 9.19 \) and the cosine distance.
Using PCA, the results for Open Set TCM-kNN are very close if the number of components used varies from 170 to 300, while using Fisherfaces the results for Open Set TCM-kNN are very close if the number of components used varies from 55 to 80. More experiments have been done randomly varying the gallery set and similar results are obtained. The threshold varies with the chosen gallery set and is determined a priori. This is different from Open Set PCA, Fisherfaces where the performance shown is obtained only if the thresholds were optimally set a posteriori using ground truth. Keeping this significant difference in mind, Open Set TCM-kNN outperforms the Open Set PCA, Fisherface classifiers. Attempts to set the thresholds ahead of time (“a prior”) for the Open Set PCA, Fisherfaces classifiers were not successful, because the intra- and inter-distance distributions for the gallery are not too powerful to characterize the behavior of the probe.

Figure 6.5: Mean Detection and Identification Rates vs. Overlap Size

Fig. 6.5 shows the mean detection and recognition rates for Open Set TCM-kNN using PCA and Fisherfaces representations with different overlay size between the gallery and probe sets. There are 150 subjects available, the size for both the gallery
and the probe sets is 75 subjects, and the overlap between the gallery list and the probe set varies from 0 to 75 subjects. We report the average results obtained over 100 randomized (over gallery and probe composition) runs. The performance goes down, almost linearly, as the overlap size increases. Fisherfaces components yield overall much better performance compared to PCA components, except for very small overlap size when the performance observed is closed but still better when using Fisherfaces than PCA components. The explanation for the observed performance is that as the size of overlap increases, it becomes more difficult to detect and identify individuals on overlap set. The performance for the Open Set PCA, Fisherfaces classifiers was very poor.

Figure 6.6: p-value Distribution of Rejection, Correct and False Recognition Using PCA (Left) and Fisherfaces (Right)

Open Set TCM-kNN also provides measures of credibility and confidence concerning the recognition decisions it makes. The p-value distribution behind Open Set TCM-kNN shows (see Fig. ??) its ability for detection and authentication [see non-overlapping correct rejection (●) and correct recognition (+) layers separated by the false recognition (○) layer] and supports the use of PSR measurement for rejection.
6.3.2 Watch List

The gallery of wanted individuals is now very small compared to the number of people expected to flood the biometric system (see Fig. 6.1b). People not on the watch list are “impostors” like, whose negative identification is sought after. The next experiment reported has 150 subjects, three images from each subject, for a total of 450 face images. We compare the Open Set PCA, Fisherfaces and Open Set TCM - kNN classifiers on small watch lists (“galleries”), whose size varies from 10 to 40 subjects, and reports the mean (average) performance (detection and identification) rates obtained over 100 randomized runs. Let the watch list size be \( n \) subjects, each of them having 2 (two) images in the gallery. Then there are \( 450 - 2n \) face images in the probe set, \( n \) stands for the number of subjects on the watch list and \( 3 \times (150 - n) \) stands for the number of face images that come from subjects that are not on the watch list. The small size of the watch list requires for stability purposes that the rejection threshold be derived from larger populations but still using as before the same statistics of intra- and inter-distance distribution for Open Set PCA, Fisherfaces and PSR distribution for Open Set TCM-kNN. The decision thresholds \( \Theta \) are derived in a manner similar to that used by cohort models in speech (see Section 6.2.1) by augmenting the gallery with different subjects randomly drawn from other FERET batches that include illumination and facial expression variation. The size of the gallery used to determine the threshold is kept constant at 80 throughout the runs so the number \( 80 - n \) of different subjects needed to augment it varies according to the size \( n \) of the watch list.

Table 6.1 and 6.2 shows the mean performance of Open Set PCA, Fisherfaces and Open Set TCM-kNN for different watch list sizes. For watch list size \( n \), the
Table 6.1: Mean Performance of Open Set PCA, Fisherfaces.

<table>
<thead>
<tr>
<th>Watch List Size</th>
<th>Eigenfaces</th>
<th>Fisherfaces</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Average Correct Reject</td>
<td>Average Correct Recognition</td>
</tr>
<tr>
<td>10</td>
<td>369.03</td>
<td>6.68</td>
</tr>
<tr>
<td>15</td>
<td>355.67</td>
<td>9.82</td>
</tr>
<tr>
<td>20</td>
<td>343.81</td>
<td>12.43</td>
</tr>
<tr>
<td>25</td>
<td>331.4</td>
<td>15.09</td>
</tr>
<tr>
<td>30</td>
<td>319.5</td>
<td>17.79</td>
</tr>
<tr>
<td>35</td>
<td>305.43</td>
<td>20.42</td>
</tr>
<tr>
<td>40</td>
<td>292.97</td>
<td>23.19</td>
</tr>
</tbody>
</table>

Accuracy (detection and identification rate) is \((\text{average correct rejection} + \text{average correct recognition})/(450 - 2n)\). The numerical results, when the number of subjects on the watch list is \(n\), should be interpreted as follows. Since the watch list size is much smaller than the number of subjects that should be rejected, the (detection and identification rate) accuracy will be very high even if all the probes are rejected. As a consequence the average correct reject number, average correct recognition number, and the detection and identification accuracy are shown for performance evaluation. The average results are better the closer the correct rejection number is to \(3 \times (150 - n)\), the closer the correct recognition number is to the watch list size, and the higher the accuracy is. Table 1 shows the average performance of Open Set PCA, Fisherfaces for different watch list sizes. The Fisherfaces components outperform PCA components for both rejection and identification decisions. As the watch list size increases, the performance drops.

The difference in performance between Fig. 6.5 and Table 6.2 indicates that the gallery size is also an important factor affecting algorithm performance. In Fig. 6.5


Table 6.2: Mean Performance of Open Set TCM-kNN.

<table>
<thead>
<tr>
<th>Watch List Size</th>
<th>Eigenfaces</th>
<th></th>
<th>Fisherfaces</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Correct</td>
<td></td>
<td>Correct</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Reject</td>
<td></td>
<td>Recognition</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Average</td>
<td></td>
<td>Average</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Accuracy</td>
<td></td>
<td>Accuracy</td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>389.74</td>
<td>7.64</td>
<td>92.41%</td>
<td>393.08</td>
</tr>
<tr>
<td>15</td>
<td>376.28</td>
<td>11.72</td>
<td>92.38%</td>
<td>380.24</td>
</tr>
<tr>
<td>20</td>
<td>364.18</td>
<td>14.12</td>
<td>92.27%</td>
<td>365.28</td>
</tr>
<tr>
<td>25</td>
<td>350.24</td>
<td>18.26</td>
<td>92.13%</td>
<td>351.84</td>
</tr>
<tr>
<td>30</td>
<td>336.94</td>
<td>21.62</td>
<td>91.94%</td>
<td>335.72</td>
</tr>
<tr>
<td>35</td>
<td>322.96</td>
<td>25.22</td>
<td>91.63%</td>
<td>322.28</td>
</tr>
<tr>
<td>40</td>
<td>309.24</td>
<td>27.98</td>
<td>91.14%</td>
<td>308.64</td>
</tr>
</tbody>
</table>

The gallery (watch list) size is always 75 subjects and only the overlap size between the gallery and probe sets changes, while in Table 6.2 the gallery size (watch list) varies according to $n$.

6.4 Discussion

This chapter expands on the Transduction Confidence Machine [42] to make it suitable for the open set recognition problem. Towards that end we introduced the Open Set TCM-kNN (Transduction Confidence Machine - k Nearest Neighbors), a novel realization of transductive inference and strangeness that is suitable for open set multi-class classification and includes a rejection option. Extensive experimental data, using challenging FERET data, shows the comparative advantages of Open Set TCM-kNN. The major contributions made include multi-class transductive inference using a priori threshold setting, effective open set identification and watch list. It can be further extended for meaningful error analysis to determine the uneven contributions to errors made by different face patterns and their use for effective data fusion, which
was shown in [47,48].

The proposed rejection functionality for open set recognition is similar to that used in detection theory, hypothesis testing, and score normalization (See Section 6.2.1). The availability of the rejection option, i.e., “none of the above” answer, in open set recognition, is similar to outlier detection and novelty detection. The comparative advantages of our proposed method come from its non-parametric implementation with strangeness and automatic threshold selection. No assumptions are made regarding the underlying probability density functions responsible for the observed data. Learning and training, driven by transduction, are local. They provide robust information to detect outlier faces, i.e., unknown faces, and to reject them accordingly. Outlier detection corresponds to change detection when faces or patterns change their appearance. The performance of open set recognition may be improved using multiple unlabeled data to learn the threshold and classifiers through co-training [12,67]. The idea of co-training is to learn two classifiers which bootstrap each other using labels for the unlabeled data [43]. Co-training leads to improved performance if at least one classifier labels at least one unlabeled instance correctly for which the other classifier currently errs. Unlabeled examples which are confidently labeled by one classifier are then added, with labels, to the training set of the other classifier.

One direction for future research concerns taking advantage of the linkage between open set recognition and part-based object representation. As demonstrated in Chapter 5, strangeness was successfully applied in part-based object recognition tasks no matter training and test images are pre-segmented or not. In order to emulate the visual learning ability of humans and deeply understand what it means to be visually similar, it will be necessary to derive open set weakly supervised object/category
learning. The key point behind this problem is to construct a valid randomness test (p-value) by integrating the strangeness of the local features, rather than the strangeness of the whole object. One possible solution is to define the strangeness of the whole object from the strangeness of local features, which satisfies Equation 2.3 and define a space where Martin-Löf randomness test can be applied validly.
Chapter 7: Conclusion and Future Work

7.1 Conclusion

This thesis presented a supervised uncertainty measure, \textit{k-Nearest Neighbor strangeness}, to score how a data point is different from others with respect to the class labels. After deeply understanding its properties, we proposed a general object recognition approach and a mechanism for open set recognition using strangeness to solve the problems of learning object or category models directly from images without segmentation and alignment and learning the rejection threshold a priori without negative training examples and modelling the distributions. It is a first step towards our long term goal of developing a generic objects/categories recognition system with rejection option, where the system can ask people as its “teacher” and acquire new knowledge on its own initiative.

For the first problem, we adopt the assumption that the object is the random constellation of local visual parts which are common across the images in the same class while discriminative from other classes. In this thesis, we considered each part independently and didn’t model their geometric relationships, which is one of my future works. Under such assumption, we showed that strangeness successfully discarded the irrelevant features, cut off the negative influence of clutter background and reduced the complexity of model learning. Representing the uncertainty of local features with respect to class labels, strangeness was further embedded in the boosting framework.
to handle the partial occlusion, local deformation, illumination and viewpoints. The corresponding local class evidence was integrated using AdaBoost and the importance of each part was provided, which may be helpful to understand what it means to be visually similar. Since strangeness is very simple and non-parametric and has a close relationship with margin (see Section 2.3), the final strong classifier has good generalization capability [76] after strangeness-based feature selection. This property was demonstrated by local recognition, object category recognition and part-based face recognition (see Chapter 5). The whole learning approach is suitable for both two-class and mult-class classification problems. It extended the recognition capability much of current object recognition methods, especially for weakly supervised object learning. It can be scaled to deal with large number of objects and/or object categories.

The strangeness was further applied to construct an Open Set TCM-kNN algorithm and a promising answer for the open set recognition problem was given. The algorithm provides a priori availability of a reject option to answer “none of the known” without modelling the distribution of any object and using negative training examples. The capability of the approach was demonstrated in a face recognition system, in particular. The comparative advantages of our proposed method come from its non-parametric implementation and automatic threshold selection. It is a productive approach for open set recognition problem with small number training examples and large number of classes.


7.2 Future work

The object recognition approach proposed in this thesis is based entirely on appearance of the local features only. The spatial relationship between features are not considered here. We hope the geometric information can be added in the boosting framework as another weak evidence and the geometric and appearance information are integrated together. To move forward from here, we need to learn how to statistically represent the geometric information, build a metric distance function and construct a valid “strangeness” measure on it. Another future research direction concerns open set weakly supervised object/category learning from local parts, which is a crucial step to emulate the visual learning ability of humans and deeply understand the visual similarity. To achieve it, we need construct a valid Martin-Löf randomness test (p-value) by integrating the strangeness of the local features.

Currently the strangeness is only defined in the supervised scenario. Unsupervised strangeness definition is still an open problem. A valid unsupervised strangeness will provide a new affinity measure, which are very useful for many vision problems. Furthermore, it is widely accepted that multiple testing examples or test sequences, such as video, can make the learning more feasible and improve the recognition performance. Incorporating temporal coherence into strangeness and p-value provides an exciting opportunity to discover new insights in learning field.
Bibliography
Bibliography


Curriculum Vitae

Fayin Li received the Bachelor of Science degree in Automatic Control from Huazhong University of Science and Technology in 1996, and the Master of Science degree from the Institute of Automation, Chinese Academy of Sciences in 1999. His research interests include computer vision, machine learning, objects/location recognition, human-computer interaction, robotics, image processing, pattern recognition, and data mining. He is also interested in time-frequency analysis and its application.

Fayin Li coauthored about 15 refereed papers, including 6 journal papers. He is the recipient of the Outstanding Graduate Student Award from school of Information Technology and Engineering, the team member of 1st place team (GMU1), ICCV 2005 vision contest.