A Fragment-Based Approach to Object Representation and Classification

Shimon Ullman, Erez Sali, and Michel Vidal-Naquet

Presented by: Guy Rapaport, M.Sc. Interdisciplinary Center Herzeliya

Agenda

- Goals
- Introduction
- Past/Related Approaches
- The Fragment-Based Approach
  - General
  - Intro to Information Theory
  - Selection of Class-Based Fragments
  - Performing Classification
  - Detecting Individual Fragments
  - Combining the Fragments and Making a Decision
- Experimental Result
- Summary

Goals

- Understanding the Fragment-Based Approach
- Introductory to Information Theory (as needed)
- Understanding the Advantages/Disadvantage and Problems

Introduction

- **Class**ification & **Identif**ication
- images of objects within a class are represented in terms of class-specific fragments
- Fragments are used as building blocks to represent large variety of images of objects in a class
- Problems:
  - Selecting the “right” set of fragments
  - Using the fragments for classification
  - Combining the fragments to take a decision

Past/Related Approaches

- Representing objects as points in higher dimensional feature space and partitioning
- Partitioning in different techniques (perceptron type algorithms, SVM…)
- Setting objects as belonging to a class – Recognition By Components
  - Using generic 3D parts (cube, cone…)
  - Object description based on their main 3D part
  - Spatial relations between parts
- Others – using fixed objects (corners, eyebrows, nose…)

The Fragment-Based Approach

- Finding and selecting the “best” fragments
  - Selecting appropriate fragments to represent a specific class
- Performing Classification of new images based on the selected fragments
  - Detection of individual fragments
  - Combining the Fragments
  - Tacking a decision

Inputs

Fragments

Selection of Class-Based Fragments

- Selecting fragments from examples
- Different fragments for different objects
- Using sets of common building blocks (fragments) to represent different object (belonging to a specific class)
- Optimized selection of the fragments (not pre-selected) according to the classification task
- Organizing the selected fragments (views. Conditions…)

Selecting Informative Fragments

- Fragments are selected using a criterion of maximizing the mutual information I(C,F) between a class C and a fragment F
- there is an a-priori probability p(C) for the appearance of an image of a given class C
- The detection of a fragment F adds information and reduces the uncertainty (measured by the entropy) of the image
- We select fragments that increase (as much as possible) the information regarding the presence of an image from the class C – in opposite way, reducing the uncertainty as much as possible – p(F|C) or p(F|NC)

Selecting Informative Fragments (cont)

- Fragment F is representative of class C if it is likely to be found in class C but not in other classes – measured by the **likelihood ratio p(F|C)/p(F|NC)**
- Fragments with high likelihood ratio are distinctive for the presence of class C – this does not necessary makes them useful fragments for representation of C – why?
- Can be distinctive but rare – generality
- … The most informative features are fragments of intermediate size

Selecting Informative Fragments (cont)

- Lets define:
  - I – Mutual Information
  - H – Entropy
- For each fragment we have
  - "Merit" – $I(C,F)=H(C)-H(C/F)$
    Measures the usefulness of a fragment F to represent a class C
  - Distinctiveness - measured by the likelihood ratio $p(F|C)/p(F|NC)$
- Both can be evaluated given $p(C)$, $p(F|C)$, $p(F|NC)$

Selecting Informative Fragments (cont)

- Fragments are selected based on their merit, and used based on their distinctiveness

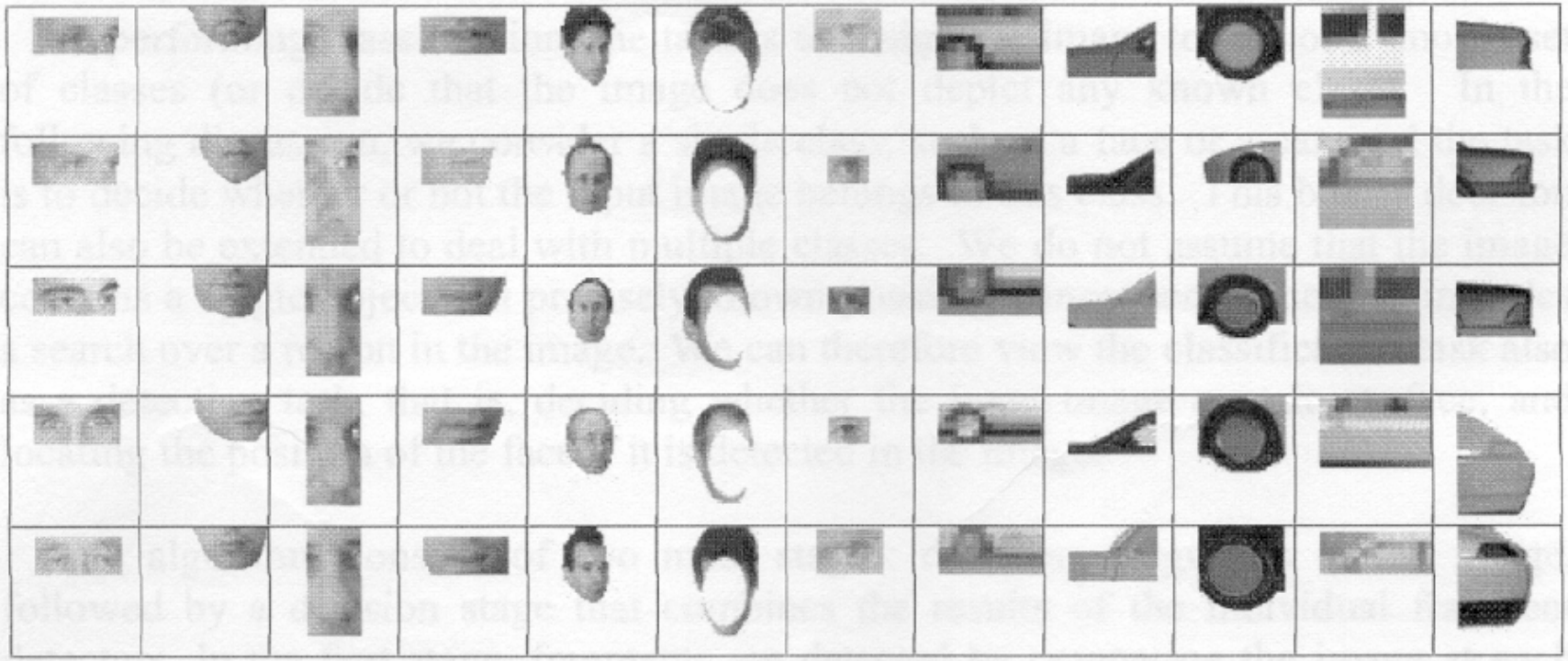


Fig. 1. Examples of face and car fragments

Selecting Informative Fragments (cont)

Selecting Fragments with High Mutual Information – the Procedure

- Given a set of images (cars) we use a pairwise comparison as an initial filter
- Performing a search of the candidate fragments in the entire database of cars and no cars database. In this manner we obtain estimations for p(F|C), p(F|NC) and assuming a particular p(C) – we can now compute the fragment's mutual information
- For each selected fragment we extend the search (for optimization) by keeping the location but changing the size and again with different resolution

Selecting Informative Fragments (cont)

- In this case they used fragments of intermediate complexity in either size or resolution

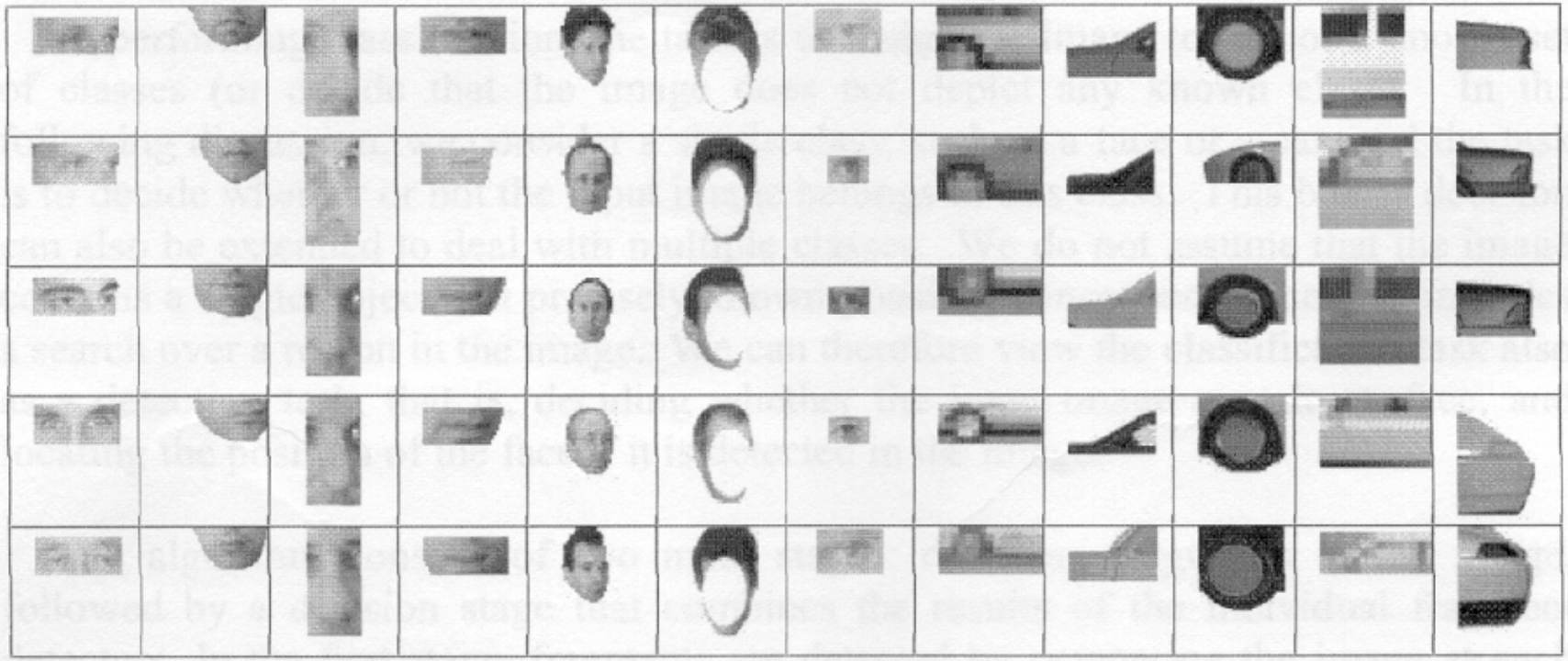


Fig. 1. Examples of face and car fragments

Performing Classification

- Performing the classification consists of two main parts:
  - Detecting fragments in the image
  - Deciding on combining the individual fragments (two different approaches:
    - A simple scheme – threshold on the amount of selected fragments
    - A complex scheme – probability distribution of the fragments

Performing Classification – Detecting Individual Fragments

- Based on direct gray-level comparison between stored fragments and input image
  - Measuring the qualitative shape similarity using the ordinal ranking of the pixels
  - Measuring the orientation difference using gradient amplitude and direction

Performing Classification – Detecting Individual Fragments

- The similarity measure D(F,H) between an image patch H and a fragment patch F is a weighted sum of their sum of these displacements *di*, the absolute orientation difference of the gradients $|\alpha_F - \alpha_H|$ and their absolute gradient difference $|G_F - G_H|$:

$$D(F,H) = k_1 \sum d_i + k_2 |\alpha_F - \alpha_H| + k_3 |G_F - G_H|$$

- mainly sensitive to the local structure of the patches and less to absolute intensity values

Performing Classification – Detecting Individual Fragments

- The Procedure:
  - Comparing local 5x5 gray level patches in each fragment to the image, using the similarity measure
  - Only regions with sufficient variability were compared (in flat-intensity regions the gradient, orientation and ordinal-order have little meaning)
  - Matching each pixel in the fragment view to the best pixel in some neighborhood around its corresponding location (allowing flexibility)
  - To detect objects at different scale in the image, the algorithm is performed on the image at several scales

Combining the Fragments and Making a Decision

- Must rely not only on the presence in the image of particular features, but also on their configurations (dependencies of features)

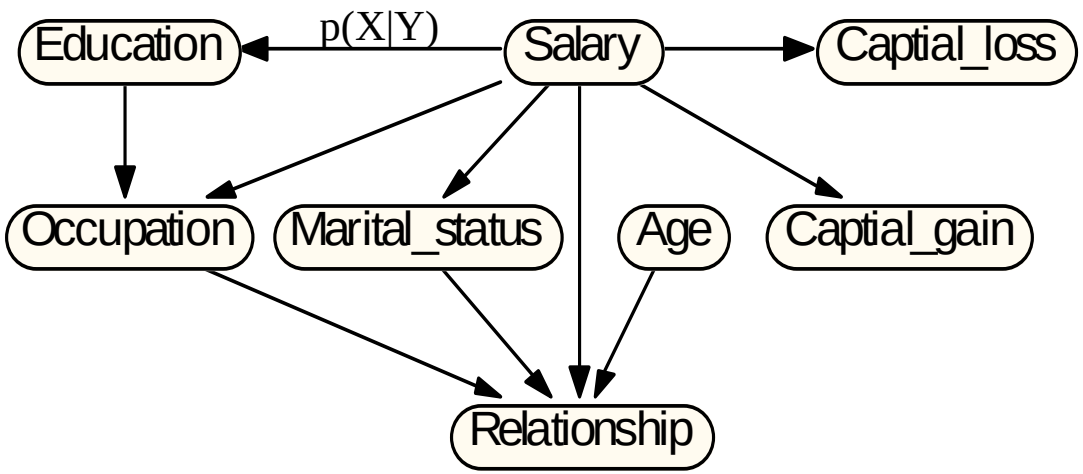
- in addition to the detection of the basic features, additional positional information, and probability distribution models of the features. In contrast, a classifier that uses more complex, class-specific visual features could employ a simpler combination scheme because the features themselves already provide good evidence about the presence of the class in question

Combining the Fragments and Making a Decision – Probability Distribution Models

- General
  - A problem of reaching a decision about the presence of a class in the images based on some set of measurements denoted by X
  - the optimal decision is obtained by evaluating the likelihood ratio defined as: $p(X|C_1)/p(X|C_0)$ where $P(X|C0)$, $P(X|C1)$ are the conditional probabilities of $\boldsymbol{X}$ within and outside the class
  - In practice, it raises computational problems
  - A popular and useful method for generating a compact representation of a probability distribution is the use of Belief-Networks, or Bayesian-Networks.

Combining the Fragments and Making a Decision – Probability Distribution Models

- directed graph where each node represents one of the variables used in the decision process
- edges correspond to dependency relationships between the variables
- parameters are conditional probabilities between inter-connected nodes



Combining the Fragments and Making a Decision – Two Combination Approaches

- **Simple approach - Naïve-Bayes**
  - The assumption that the entries of the feature vector can be considered independent when computing likelihood ratio $p(X|C_1)/p(X|C_0)$
  - In practice, it means we first measure the probability of each fragment Xi within and outside the class
  - To reach a decision we simply multiply the relevant probabilities together
  - This method assumes independence between the different fragments

Combining the Fragments and Making a Decision – Two Combination Approaches

- The actual computation in our classification scheme was performed using the fragment types, rather than the fragments themselves
- for each fragment type (such as a hairline or eye region), the best-matching fragment was selected.

Combining the Fragments and Making a Decision – Two Combination Approaches

- **Dependence-Tree Combination**
  - A simple Bayesian-Network describing a probability distribution that incorporates relevant pairwise dependencies between variables
  - Features are organized in a tree structure that represents statistical dependencies
  - The tree structure permits the use of some, but not all, of the dependencies between features
  - An optimal tree representation is constructed from information regarding pairwise dependencies in the data

Combining the Fragments and Making a Decision – Two Combination Approaches

- The probability of an input vector is computed by multiplying together the probabilities of each node given the value of its parent. More formally:

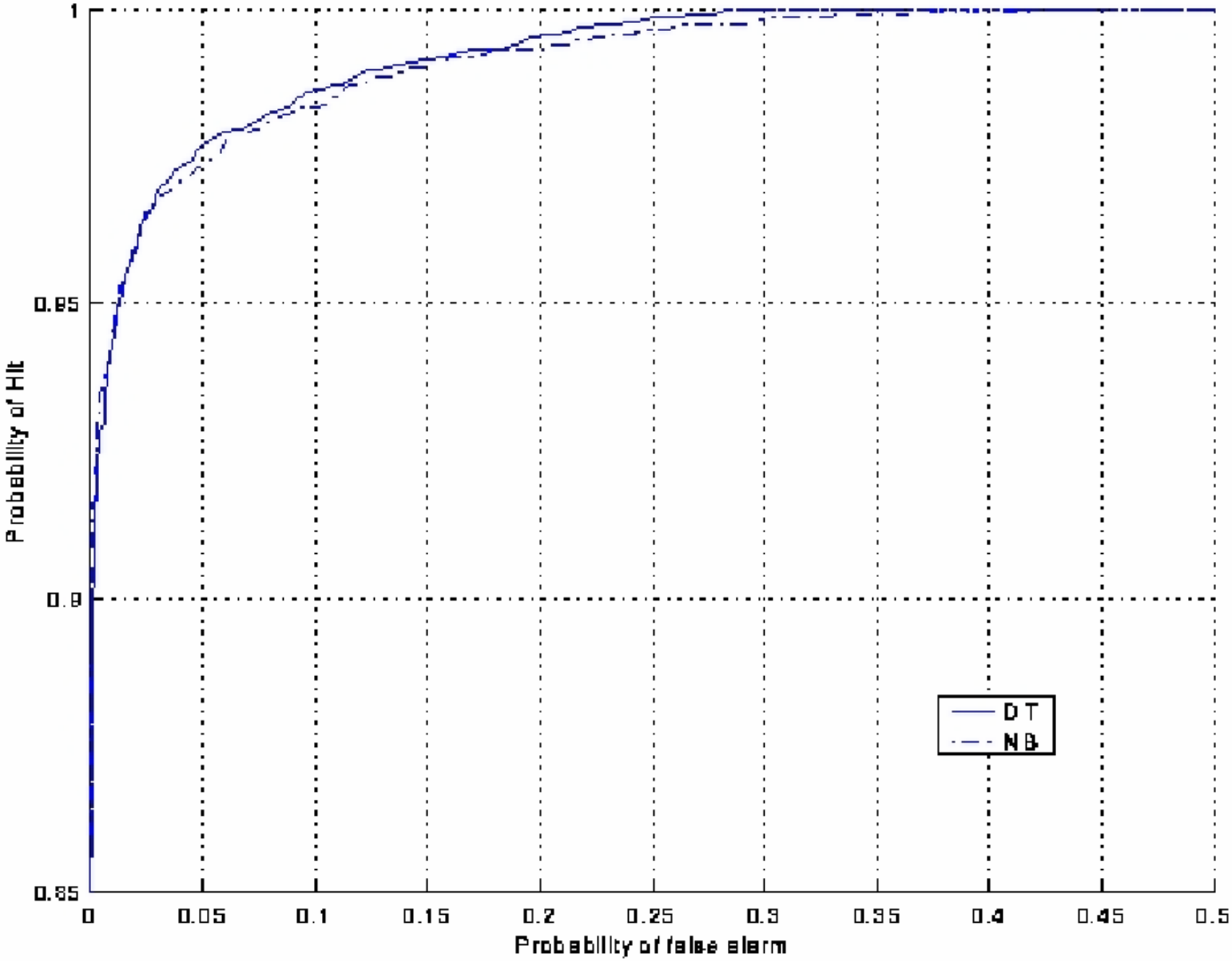
$$P(X_1,\ldots X_n|C) = P(X_1|C) * \prod_{I=2}^{N} P(X_i|X_{j(i)},C)$$

- j(i) represents the parent of node I
- X1 is the root of the tree (which represent the class variable – no parent)

Trade-Off between Feature Complexity and Combination Complexity

- Implementation of the two mentioned combination schemes
- Comparing a simple combination scheme, based primarily on the presence or absence of fragments in the image, and a more elaborate scheme that uses a refined model of the probability distribution of the fragments

Receiver Operating Characteristic (ROC) curves for both classifiers



Trade-Off between Feature Complexity and Combination Complexity

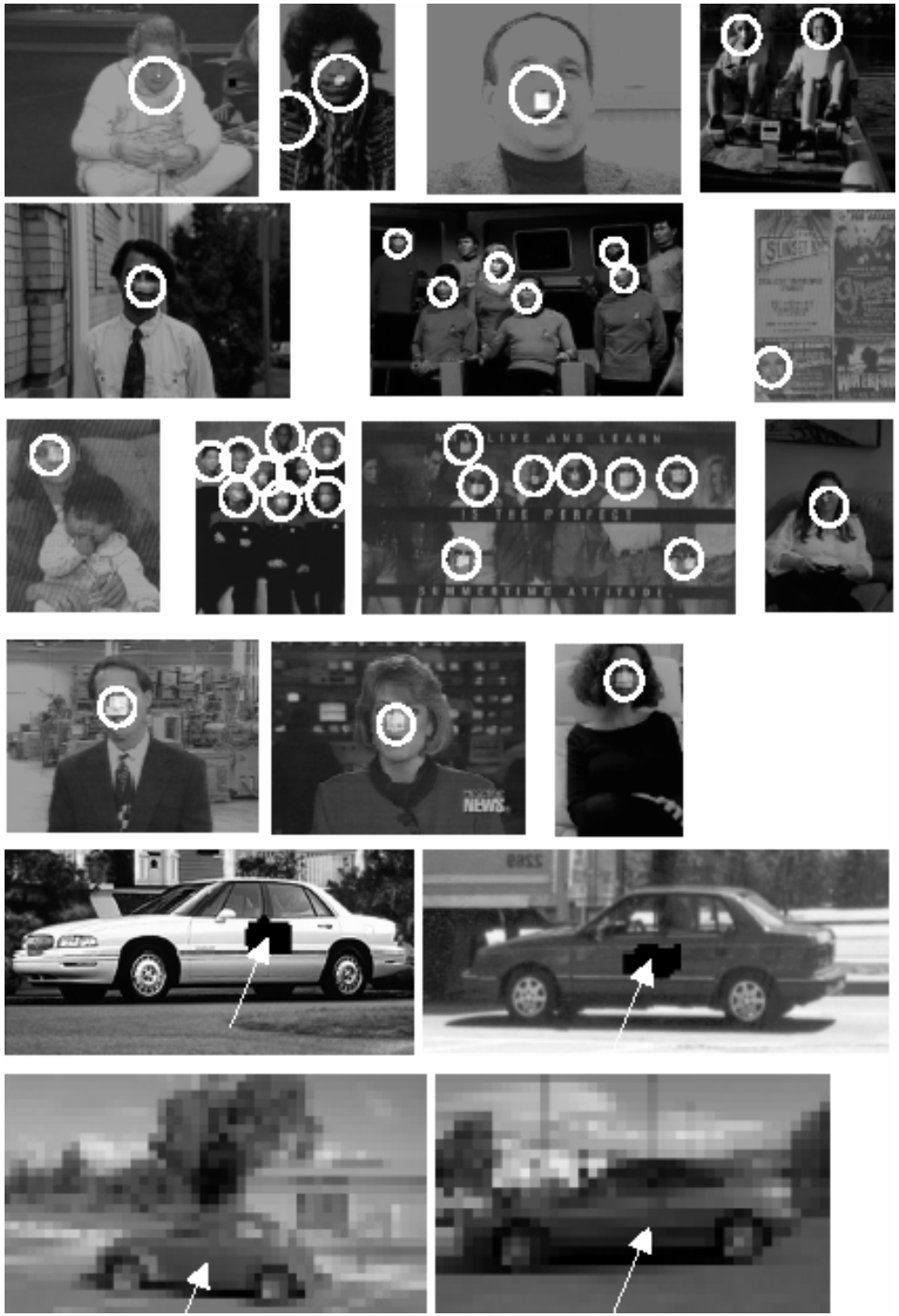
- The curves for both methods are almost identical
- This property of the classifier depends on the features used for classification
- When simple generic features are used, the dependencies between features at different locations are important
- When complex features are used, as in this case, then a simpler combination scheme will suffice

Experimental Results

- Tested on face and car views
- Numerous runs with different databases
- Different details of the fragment extraction and combination procedures
- In a specific run - tested face detection, using a set of 1104 part views,taken from a set of 23 male face views under three illuminations and three horizontal rotations
- Fragments were grouped into eight fragment types – eye pair, nose, mouth, forehead, low-resolution view, mouth and chin, single eye and face outline
- For cars, we used 153 parts of 6 types

Experimental Results

- Illumination
- Views
- Numerous objects
- Male & Female



Conclusions

- This method generalizes well to new objects within the class of interest
- Small set of examples needed
- Various conditions
- Real images and drawings
- Generalization
- Maintaining low false alarm rates on images that did not contain the object

Conclusions (cont)

- Capable of obtaining significant position invariance, without using explicit representation of the spatial relationships between fragments
- The insensitivity to position as well as to other viewing parameters was obtained primarily by the use of a redundant set of overlapping fragments, including fragments of intermediate size and higher resolution, and fragments of larger size and lower resolution

Thank You

Introductory to

Information Theory

Basic Concept

- What is information?
  - Attneave (1959): Information is that which removes or reduces uncertainty
- How to measure uncertainty?
  - A quantitative measure of uncertainty should have at least the following properties
    - If the outcome of an event can be predicted with a 100% accuracy, then the uncertainty of an event is zero
    - The uncertainty of an event increases with the number of possible outcomes
    - For the same number of outcomes, the uncertainty is maximal if each outcome has the same probability

Basic Concept

- **Notation:**
  - capital letters - discrete random variable
  - lowercase letters - particular value of a variable
  - Pr←X=x) or Pr←x) - probability that X takes on the particular value x
  - Pr←X= x,Y= y) or Pr←x,y) - The probability that X= $x$and Y= y–*joint probability*
  - Pr←X= x♣ Y= y) or Pr←x♣ y)› conditional probability that X= $x$given Y=

Basic Concept

- Fair coin – {Pr(Coin = Heads) = 1/2; Pr(Coin = Tails) = ½}
- Biased coin – {Pr(BiasedCoin = Heads) = 0.9; Pr(BiasedCoin = Tails) = 0.1}
- More doubts regarding the fair coin
- Can we make the notion of uncertainty or doubt quantitative?

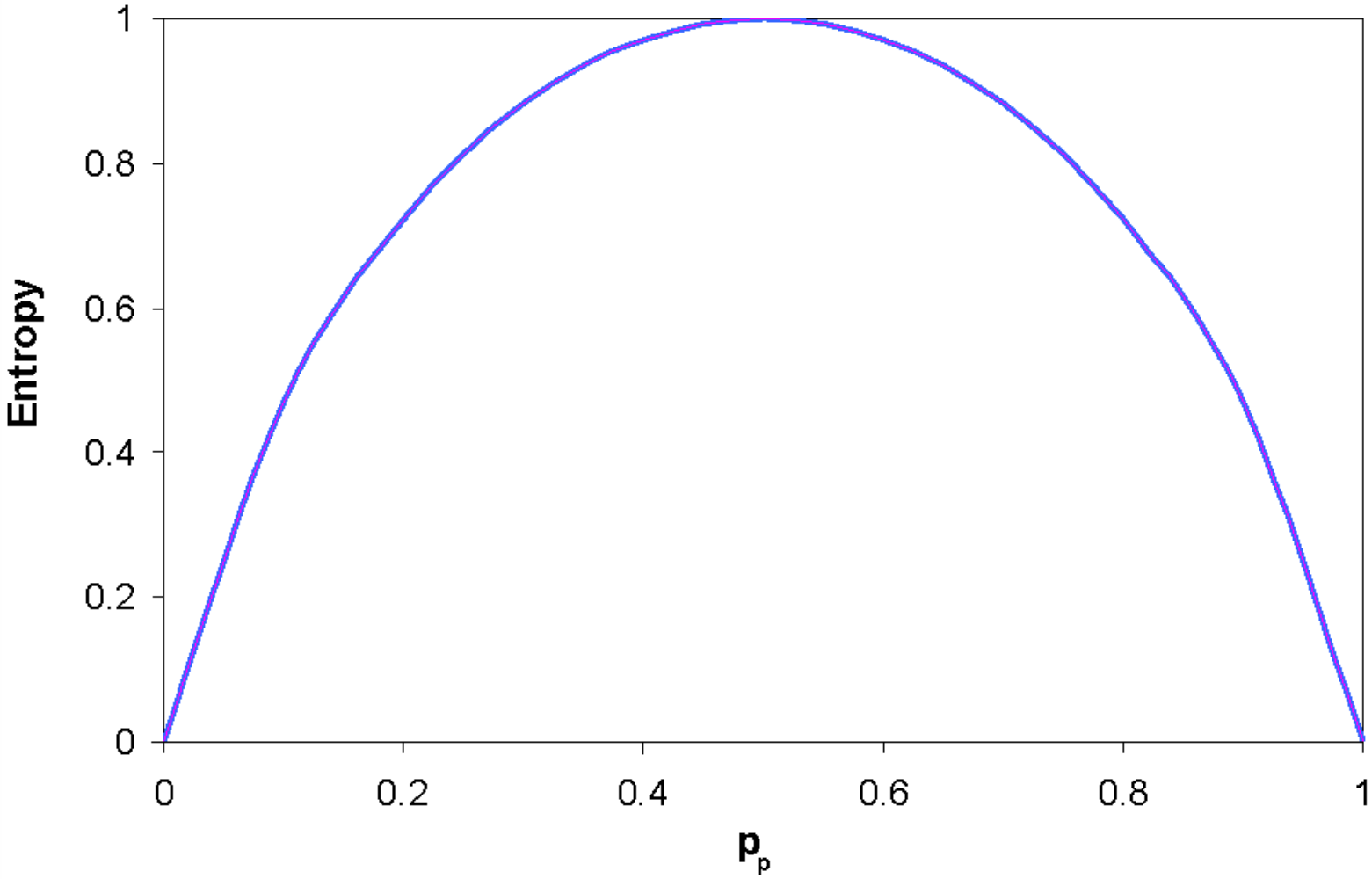
Making the Uncertainty Quantities

- $Features for the measure.$
  - H to be maximized by a uniform distribution
  - H is a continuous function of the probabilities
    - An arbitrarily small change in the probabilities should lead to an arbitrarily small change in H
  - H to be a function of the distribution itself and not a function of how we group events within that distribution

Entropy

- **Entropy** $H(X) = -\sum p(x)\log p(x)$
  Where:
  X - a discrete random variable
  x - value of X
  p (x) - probability of x
- **Interpretation**: measure of uncertainty of X
- Properties of Entropy:
  - H(X)=0 if and only if the outcome is deterministic (all p(x) but one are zero)
  - $H(X) \leq \log[\text{number of outcomes}]$, H(X) is a maximum, when all outcomes are equiprobable
  - If all the outcomes have the same probability, then H(X) is a monotonic increasing function of the number of outcomes

A Bit More



$$-H(X|Y) = -\sum_{x}\sum_{y} p(x,y)\log p(x|y)$$

Entropy and not Variance

- Entropy gives the same results regardless for example of different peaks in the distribution
  - Peaks are adjacent (close to each other) = lower variance and vice versa
- Entropy is a function of the distribution itself and not a function of how we group events within that distribution

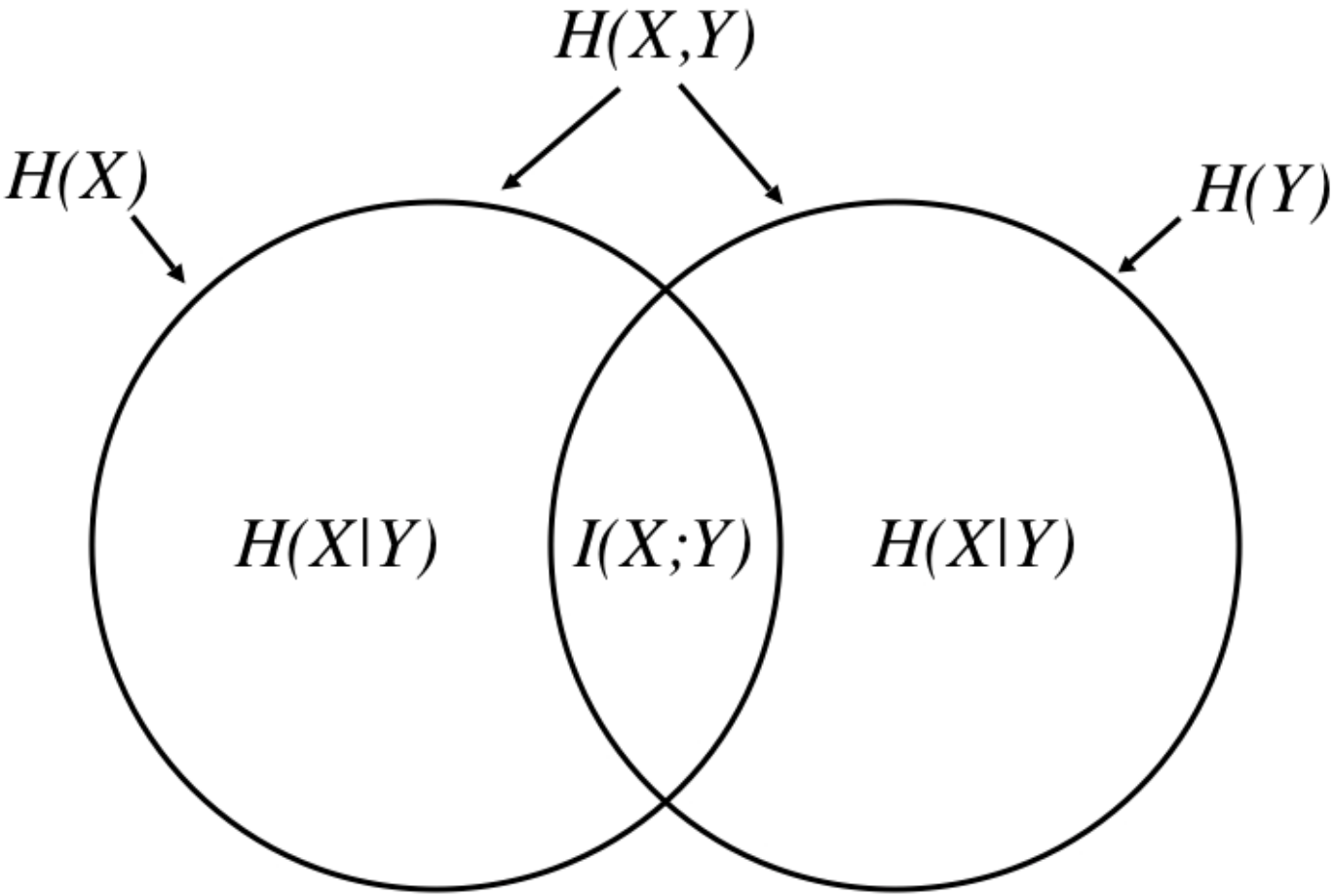
Mutual Information

- We define the mutual information, I[X; Y] of two random variables $X$and Yas.

$$I[X;Y] \equiv \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} \Pr(x,y) \log_2 \left[\frac{\Pr(x,y)}{\Pr(x)\Pr(y)} \right]$$

- reduction in uncertainty of one variable due to knowledge of another - Ycarries information about X

All Together



Venn diagram of relations between 2 variable entropies

Kullback-Leibler Distance

- a quantity which measures the difference between two probability distributions

$$KL(p,q) = \sum_x p(x) \log \frac{p(x)}{q(x)}$$

- This distance is zero when *P(X,Y)=P(X)P(Y),* i.e., when *X* and *Y* are independent.