Computer Vision

MSc Informatics option GVR
James L. Crowley

Fall Semester 12 Novembre 2009

Lesson 6

View Invariant Bayesian Recognition

Lesson Outline:

1 Scale and Rotation Invariant Image Description

1.1 Image Scale Space:

Continuous Case.

Let P(x,y) be the image.
Let G(x, y, 2^s) by a Gaussian function of scale $\sigma=2^s$

Continuous x, y, s: $P(x, y, s) = P(x,y)* G(x, y, 2^s)$

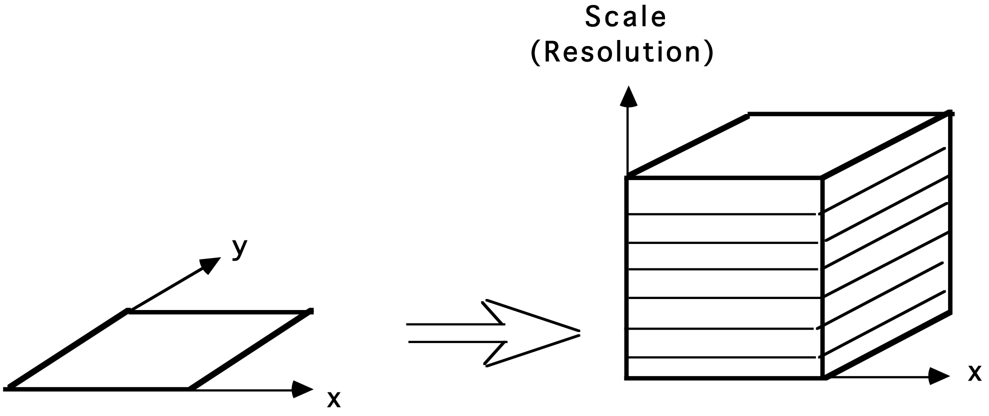
Note that the scale axis (s) is logarithmic. $s = \log(2^s)$

Intuitive consequence: If a shape in an image is made larger by $B = 2^d$

$p(x,y) \text{ -> } p(x2^d, y2^d)$

Then the scale space projection is shifted by s

$P(x,y,s+d) = p(x2^d, y2^d) * G(x, y, 2^s)$



The appearance of a pattern in the image results in a unique structure in P(x, y, s).
This structure is "equivariant" in position, scale and rotation.
Translate the pattern by Δx, Δy and the structure translates by Δx, Δy in P(x, y, s).

Rotate by θ in x,y and the structure rotates by θ in P(x, y, s).

Scale by a factor of 2^s, and the structure translates by s in P(x, y, s).

Scale space :
Separates global structure from fine detail.
Provides context for recognition.
Provides a description that is invariant to position, orientation and scale.

1.2 Discrete Scale Space - Scale invariant impulse response.

In a computer, we need to discretize x, y, and s.

Let P(i,j) be a discrete representation for $P(x,y) = P(i\Delta x, j\Delta x)$
Suppose P(i,j) is an image array of size M x M pixels.

We propose to sample scale with a step size of $\Delta\sigma = 2^{1/2}$ so that $\sigma_k=2^{k/2}$

Note that scale space "dilates" the Gaussian impulse response by 2^s.

$$P(x,y,s) = P(x,y)* G(x, y, 2^s)$$

As the Gaussian impulse response dilates, the sample density can also dilate.

$$p(i\ \Delta x_s, j\ \Delta x_s, \Delta s) \quad \text{such that } \Delta x_k = 2^{k/2}$$

$$P(x,y,2^s) = , \quad \text{where } \Delta x_k = 2^{k/2}$$

For a Gaussian Kernel filter $G(i,j,k) = G(x, y, \sigma_k=2^{k/2})$

The image pyramid becomes :

$$P(i,j,k) = p(i\ 2^{k/2}, j\ 2^{k/2}, 2^{k/2}) = P(i,j) * G(i\ 2^{k/2}, j\ 2^{k/2}, 2^{k/2})$$

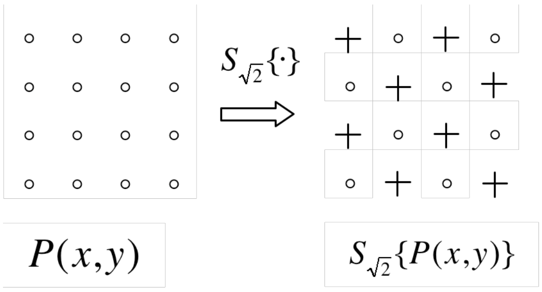
for $1 \leq k \leq M-4$

1.3 Diagonal, Square root of two Sampling

Problem : How can we sample an image for for odd k? $\Delta x = 2^{k/2} = 2^{(k-1)/2}\sqrt{2}$

for k odd, $\Delta x_k = \{1, 2, 4, 8\ldots\}$
for k even, $\Delta x_k = \{\sqrt{2}, 2\sqrt{2}, 4\sqrt{2}, 8\sqrt{2}, \ldots\}$

How ? with the diagonal sampling operator $S_{\sqrt{2}}\{\}$

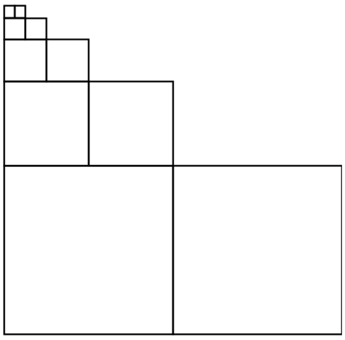


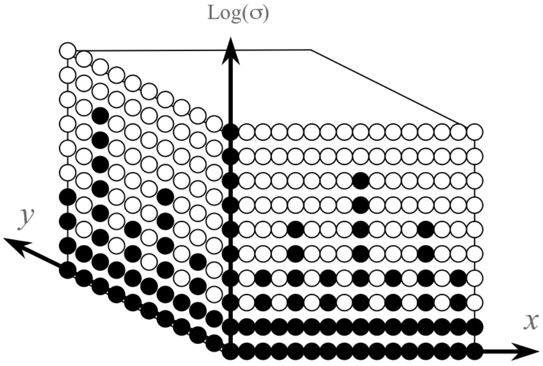
For k even, the √2 resampling operator, $S_{\sqrt{2}}^{k}\{\}$, selects even columns of even rows and odd columns of odd rows.

For k odd, diagonal sample operator eliminates every second column (starting with even columns on even rows and odd columns on odd rows). For k odd, resampling eliminates every second row (odd rows).

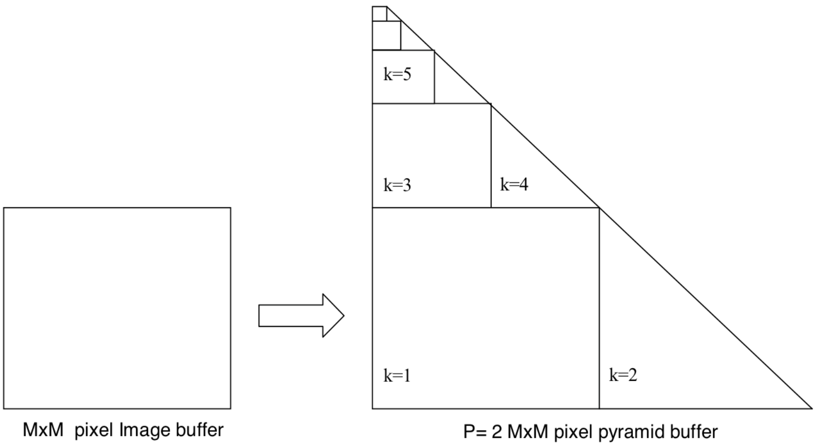
$$S_{\sqrt{2}^k}\{P\{x,y)\} = \begin{cases} P(x,y) & \text{if } (x+y)^2 \text{ Mod } 2^{k-1} = 0 \\ 0 & \text{otherwise} \end{cases}$$

Data Structure





The even numbered images are diagonally sampled, eliminating half the pixels.



For an image of size MxM, number of pixels is

$$P = MxM \text{ x } (1 + ½ + ¼ + \ldots) = 2M^2$$

Within such a structure, the derivatives can be approximated as differences:

$$P_x(i,j,k) =< P(i,j), G_x(i,j,2^{k/2}) \approx P(i+1,j,k) - P(i-1,j,k)$$

$$P_y(i,j,k) =< P(i,j), G_x(i,j,2^{k/2}) \approx P(i,j+1,k) - P(i,j-1,k)$$

$$P_{xx}(i,j,k) =< P(i,j), G_{xx}(i,j,2^{k/2}) \approx P(i+1,j,k) - 2P(i,j,k) + P(i-1,j,k)$$

$$P_{yy}(i,j,k) =< P(i,j), G_{yy}(i,j,2^{k/2}) \approx P(i,j+1,k) - 2P(i,j,k) + P(i,j-1,k)$$

$$P_{xy}(i,j,k) =< P(i,j), G_{xy}(i,j,2^{k/2}) >$$
$$\approx P(i+1,j+1,k) - P(i-1,j+1,k) - P(i+1,j-1,k) + P(i-1,j-1,k)$$

Diffusion Equation: $\nabla^2 G_x(i,j,\sigma) = G_{xx}(i,j,\sigma) + G_{yy}(i,j,\sigma) = \dfrac{\partial G(i,j,\sigma)}{\partial \sigma}$

As a consequence: $\nabla^2 G(i,j,\sigma) \approx G(i,j,\sigma_1) - G(i,j,\sigma_2)$

This typically requires $\sigma_1 \geq \sqrt{2}\ \sigma_2$

Thus it is common to use:

$$\nabla^2 P(i,j,k) = <p(i,j), \nabla^2 G(i,j,\sigma_k)> \approx P(i,j,k) - P(i,j,k-1)$$

1.4 Cascade Convolution Pyramid Algorithm:

Cost of computing p(i,j,k) is

$$C= O(M^2((N_0+1)^2+(N_1+1)^2+(N_2+1)^2+\ldots+(N_{M-4}+1)^2))$$

if we use "seperable" convolution:

$$P(i,j) * G(i,j, 2^{k/2}) = P(i,j) * G(i, 2^{k/2}) * G(j, 2^{k/2})$$

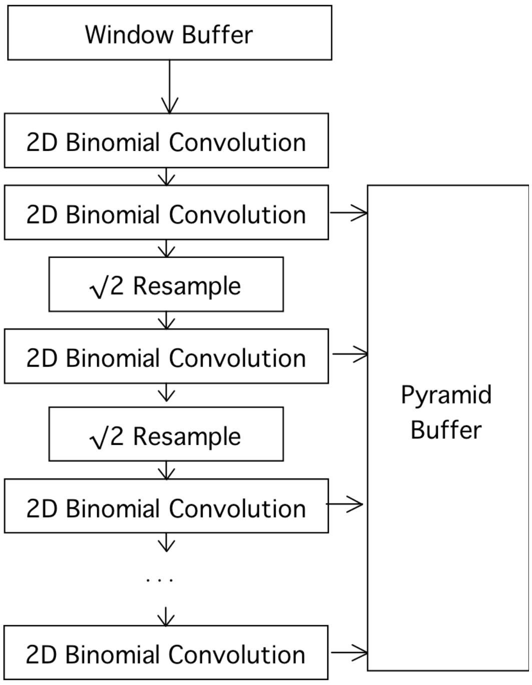
then

$$C= O(M^2 \cdot 2(N_0+N_1+N_2+N_3+\ldots+N_{M-4}+M-4+1)$$

$$C= O(M^2 \cdot 2(8+16+32+64+\ldots+ N_{M-4})+6).$$

Practically, the computational cost is exorbitant.

We can use Cascade Convolution Methods to reduce cost to O(N)



1.5 Scale Invariant Interest Points

Maximal points in the image derivatives provide keypoints.
In an image scale space, these points are scale invariant.

Example: maxima in the lapacian as invariant "keypoints"
(often called "interest points").

Recall the Laplacian of the image :

$$\nabla^2 P(x,y,s) = P * \nabla^2 G(x,y,\sigma) = P * G_{xx}(x,y,\sigma) + P * G_{yy}(x,y,\sigma) \approx P * \nabla^2 G(x,y,\sigma_1) - P * \nabla^2 G(x,y,\sigma_2)$$

Scale invariant keypoints are given by

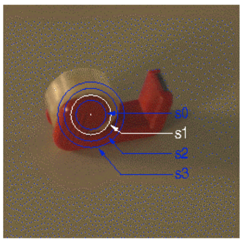
$$(\mathrm{x,y,s}) = \underset{x,y,s}{\arg-\max}\{\nabla^2 P(x,y,s)\}$$

Since $\nabla^2 \mathrm{P(i,j,k)} = \langle \mathrm{P(i,j)}, \nabla^2 \mathrm{G(i,j,}\sigma_k) \rangle \approx \mathrm{P(i,j,k)} - \mathrm{P(i,j, k{-}1)}$

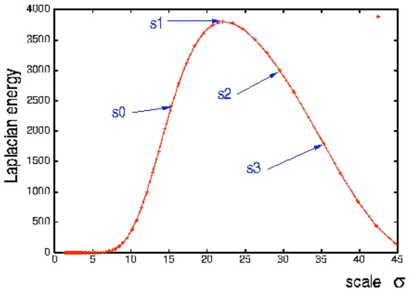
We can detect scale invariant keypoints as

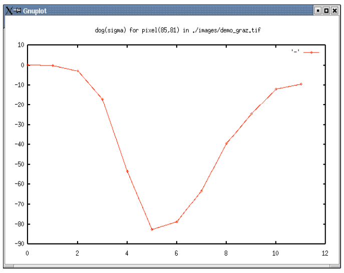
$$(\mathrm{i,j,k})_n = \underset{i,j,k}{\arg-\max}\{\nabla^2 P(i,j,k)\}$$

Examples:



zero crossing of Laplacian at si





Maximally stable invariant points are found as :

$$X(i,j,k) = \arg\text{-}\max_{i,j,k}\{P(i,j,k) - P(i,j,k-1)\}$$

Such points are used for tracking, for image registration, and as feature points for recognition.

In fact, the scale of the maximal laplacians is an invariant at ALL image points.

The scale σ_i is an "invariant" for the appearance at P(i,j).

$$\sigma_i = \arg\text{-}\max_{\sigma}\{P * \nabla^2 G(i,j,\sigma)\}$$
$$\sigma_i = \arg\text{-}\max_{\sigma}\{\nabla^2_{\sigma=2^k} P(i,j)\}$$
$$\sigma_i = \arg\text{-}\max_{k}\{P(i,j,k) - P(i,j,k-1)\}$$

1.6 Scale Invariant Feature Transform (SIFT)

Notes not yet available

2 Bayesian Recognition

Recognition is a fundamental ability for intelligence, and indeed for all life.
To survive, any creature must be able to recognize food, enemies and friends.

Recognition: The fact to recognize, to identify an object as itself.
Identify: To recognize an entity as an individual
Classify: The recognize an individual as a member of a class.

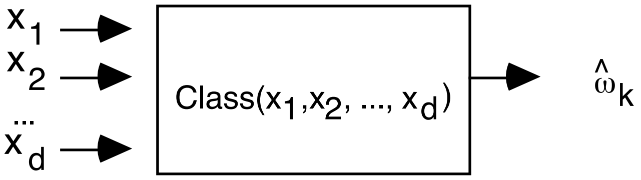
A class is defined by a membership test.

Classification is a process of associated an entity (or an event) as a member of a class. The entity is described by a vector of features, provided by an observation. The assignment of an entity to a class provided by a test made on the feature vector.

Features: observable properties that permit discrimination between classes.
A set of D features, x_d, are assembled into a feature vector $\vec{X}$

$$\vec{X} = \begin{pmatrix} x_1 \\ x_2 \\ \dots \\ x_D \end{pmatrix}$$

For a feature vector, $\vec{X}$, a classifier is a process that proposes the identity of the class. This arrives in the form of a proposition $\hat{\omega}_k = E \in$ Class C_k



The techniques from Pattern recognition and statistics provide a variety of methods to construct membership tests for classification of observations. The most appropriate technique depends on the number and nature of the classes and the features.

There are two families of technique: Generative and discriminative.
These correspond to the two methods to define a set:

Extension: Provide a list of members.
Intension: Provide a conjunction of predicates.

Generative methods compare the pattern to a set of prototype examples.
Discriminative methods apply a set of tests.

In either case, our objective is to minimize the probability of error.

$$\hat{\omega}_k = \underset{k}{\arg-\max}\left\{\Pr(E \in T_k \mid \vec{X}\right\}$$

The operator "|" is called to as "given" or "provided that". It is the Bayesian conditional operator.

For a Generative method, we enumerate the M examples of the K classes. The estimate is the most similar, as provided by some simarity function.
Thus for an observed unknown even X, the "estimated" class, $\omega^{\wedge}_k$ is given by :

$$\hat{\omega}_k = \forall_k \forall_m : \underset{k}{\arg-\max}\left\{Sim(\vec{X}, \vec{X}^k_m)\right\}$$

Simple Euclidean distance is often used to measure similarity.

$$\hat{\omega}_k = \forall_k \forall_m : \underset{k}{\arg-\max}\left\{\| \vec{X}, \vec{X}^k_m \|\right\}$$

A more intelligent method is to no normalize distance by a Metric Λ_λ

$$\hat{\omega}_k = \forall_k \forall_m : \underset{k}{\arg-\max}\left\{(\vec{X} - \vec{X}^k_m)^T \Lambda_k (\vec{X} \;\; - \vec{X}^k_m) \;\right\}$$

We can avoid having to scan all samples by replacing samples of the same class with the average of the samples.

$$\vec{\mu}_k = E\{\vec{X}^k_m\} = \frac{1}{M}\sum_{m=1}^{M} \vec{X}^k_m$$

Then:

$$\hat{\omega}_k = \forall_k : \underset{k}{\arg-\max}\left\{(\vec{X} - \vec{\mu}_k)^T \Lambda_k (\vec{X} - \vec{\mu}_k)\right\}$$

where the metric Λ_k is provide by the inverse of the class covariance :$\Lambda_k = C_k^{-1}$

$$C_k = E\{(\vec{X}^k_m - \vec{\mu}_k)^2\}$$

Discriminative tests avoid iterating through the M examples of each class by compiling a series of simple tests. These can be combined in a variety of ways.

A classical (and effective) means is by vote over as large set of simple linear classification functions. We will see more of this later.

2.1 Bayesian Classification

With a Baysian approach, the tests are designed to minimize the number of errors.
False positives and false negatives count equally as errors.
An alternative would be to include the cost of error, which may not be the same for a false positive and a false negative. This is an easy extension.

Let ω_k be the proposition that the event belongs to class k: $\omega_k = E \in T_k$

ω_k	Proposition that event E $\in$ the class k
$p(\omega_k) = p(E \in T_k)$	Probability that E is a member of class k

Given an observation $\vec{X}$, the decision criteria is

$$p(\omega_k \mid \vec{X}) = \Pr(E \in T_k \text{ given } \vec{X})$$

$$\hat{\omega}_k = \underset{\omega_k}{\text{arg-max}} \{ p(\omega_k \mid \vec{X}) \}$$

The meaning of "given" is provided by Bayes Rule:

$$p(\omega_k \mid \vec{X}) = \frac{p(\vec{X} \mid \omega_k) p(\omega_k)}{p(\vec{X})}$$

2.2 The probability of an event

There are two ways to define "probability": 1) Statistics : Frequency of Occurrence: The fraction of times that something is true, or 2) Probability: Using a systems of axioms.

A Frequency based, or statistical approach is more intuitive, but not always possible to apply. In some cases, the axioms of probability theory can provide a solution that is not possible from frequency of occurrence.

In either case, probability is a function that returns a number between 0 and 1.
$\Pr() \in [0, 1]$.

2.2.1 Probabilty as Frequency of Occurence.

A frequency based definition of probability is sufficient for many practical vision problems.

Given M observations of random events, of which M_k belong to the class k.
The probability of observing an event E of class k is

$$p(E \in A_k) \equiv \lim_{M \to \infty} \left\{ \frac{M_k}{M} \right\}$$

For the practical case where M is finite, $p(E \in \text{class } k) \approx \frac{M_k}{M}$
The precision of this approximation depends on the number of sample, M.

2.2.2 Axiomatic Definition

An axiomatic definition makes it possible to apply analytical techniques to the design of classification systems. Only three postulates (or axioms) are necessary:
In the following, let E be an event, let S be the set of all events, and let A_k be set of events that belong to class k.

Postulate 1 : $\forall A_k \in S$: $p(E \in A_k) \geq 0$
Postulate 2 : $p(E \in S) = 1$
Postulate 3 :
$\forall A_i, A_j \in S$ such that $A_i \cap A_j = \varnothing$: $p(E \in A_i \cup A_j) = p(E \in A_i) + p(E \in A_j)$

2.3 The probability of the value of a random Variable.

For integer x, such that $x \in [x_{min}, x_{max}]$, we can consider each value of x as a class.

We can then estimate the probability for each class using M observations $\{X_m\}$.

To estimate the probability of a value, we count the number of times it occurs.
For this we use a table of "frequency of occurrence", also known as a "histogram", h(x).

The existence of computers with gigabytes of memory has made the computation of such tables practical.

We use the table to count the number of times each value occurs:

$$\forall m=1, M \ : \ h(X_m) := h(X_m) + 1; \ \ M := M+1;$$

Thus, the probability of a value, $X \in [X_{min}, X_{max}]$ is the frequency of occurrence of the value.

The probability that a random value X takes a given value x is

$$p(X=x) = \frac{1}{M} \ h(x)$$

Problem: How many observations, M, do we need?

Answer:

Given N possible values of x, and M observations, in the worst case:

"average error" is proportional to O(N/M).

Rule of thumb. For most applications, we need M = 10N (10 samples per "cell").

2.4 Bayes Rule

Let E represent a random event from an event "generator" (for example a sensor).

Consider 2 independent classes of events A and B such that

E may be $A \cap B$ or $\neg A \cap B$ or $A \cap \neg B$ or $\neg A \cap \neg B$

For an E we can form 2 propositions, p and q.

$p \equiv E \in A$ et $q \equiv E \in B$

thus the probability of each proposition is

$P(p) \equiv \Pr\{E \in A\}$ and
$P(q) \equiv \Pr\{E \in B\}$.

From the axioms of probability :

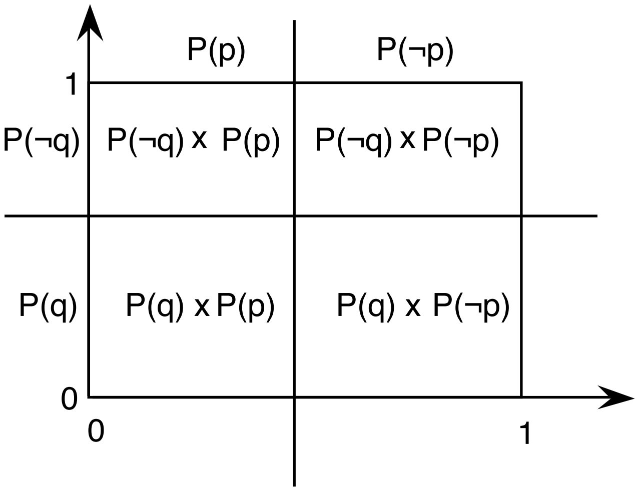
$P(q) + P(\neg q) = 1$.

$P(p \wedge q)$ is the joint probability of p and q.

IF A and B are independent then p and q are independent :

$P(p \wedge q) = P(p) \cdot P(q)$,
$P(p \vee q) = P(p) + P(q)$.

Graphically this is



	P(p)	P(¬p)
P(¬q)	P(¬q) x P(p)	P(¬q) x P(¬p)
P(q)	P(q) x P(p)	P(q) x P(¬p)

$$P(p \wedge q) + P(p \wedge \neg q) + P(\neg p \wedge q) + P(\neg p \wedge \neg q) = 1$$

The marginal probabilities are:

$$P(p) = P(p \wedge q) + P(p \wedge \neg q)$$
$$P(q) = P(p \wedge q) + P(\neg p \wedge q)$$

The "conditional" probabilities are defined as :

$$P(q \mid p) = \frac{P(p \wedge q)}{P(p)} = \frac{P(p \wedge q)}{P(p \wedge q) + P(p \wedge \neg q)}$$

and

$$P(p \mid q) = \frac{P(p \wedge q)}{P(q)} = \frac{P(p \wedge q)}{P(p \wedge q) + P(\neg p \wedge q)}$$

that is, the probability that p is true, given that q is true is P(p|q).

By algebra :

$$P(q \mid p)\, P(p) = P(p \wedge q) = P(p \mid q)\, P(q)$$

thus

$$P(q \mid p)\, P(p) = P(p \mid q)\, P(q)$$

This is Bayes Rule. I can also be written:

$$P(q \mid p) = \frac{P(p \mid q)\, P(q)}{P(p)}$$

P(q | p) is the conditional or the "posterior" probability of q.

3 Classification by Ratio of Histograms of pixel values

Histograms provide an alternate view of Bayes Rule.

3.1 Histograms

As we saw, for integer x from a bounded set of values, such that $x \in [x_{min}, x_{max}]$,

the probability that a random observation X takes on x is

$$P(X=x) = \frac{1}{M} h(x)$$

The validity of this depends on the ratio of the number of sample observations M and the number of cells in the histogram Q=N

This is true for vectors as well as values.

For a vector of D values $\vec{x}$ the table has D dimensions. $h(x_1, x_2, \ldots, x_D) = h(\vec{x})$

The average error depends on the ration $Q=N^D$ and M. : $E_{ms} \sim O(\frac{Q}{M})$

We need to assure that $M >> Q = N^d$

As a general rule : $M = 10N^d$

3.2 Example: Object detection by pigment color

We can use Bayes rule to detect objects based on their pigment.

The observed chrominence $C = \begin{pmatrix} c_1 \\ c_2 \end{pmatrix}$ is a signature for an object.

$$c_1 = r = \frac{R}{R+V+B} \qquad c_2 = v = \frac{V}{R+V+B}$$

Suppose that these are coded with N values between 0 and N – 1

$$c_1 = \text{Round}\left((N-1) \cdot \frac{R}{R+G+B} \right) \quad c_2 = \text{Round}\left((N-1) \cdot \frac{G}{R+G+B} \right)$$

Allocate a 2D table $h(c_1, c_2)$., of size N x N.

(for example, for 32 x 32 Q = 32 x 32 = 1024 cellules)

For each pixel in the image $\vec{C} = C(i, j)$

$$h(\vec{C}) := h(\vec{C}) + 1$$

That is $h(c_1, c_2) := h(c_1, c_2) + 1$

After M pixels, the chrominance histogram $h(\vec{C})$, gives :

$$P(\vec{C}) \approx \frac{1}{M} h(\vec{C})$$

Consider a region W of M_o pixels of a known object class O.

$$\forall(i,j) \in W \ : \ h_o(\vec{C}(i,j)) := h_o(\vec{C}(i,j)) + 1$$

Then $\vec{C}(i, j) = \begin{pmatrix} r \\ v \end{pmatrix}(i, j)$: $p(\vec{C} \mid \text{objet}) \approx \frac{1}{M_o} h_o(\vec{C})$

Because W is part of the image, the probability of observing a pixel from W is

$$P(W) = \frac{M_o}{M}$$

From Bayes rule, for any pixel $\vec{C}(i, j)$ the probability that it belongs to O is

$$p(\text{objet} \mid \vec{C})$$

for S images de IxJ pixels we have $M=S \cdot I \cdot J$ pixels.

Suppose that each contains a known region of the object, W_s. so that
we have M_o total pixels of the object in the S images.

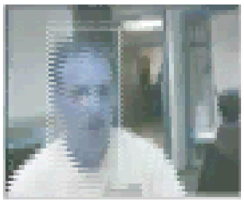
$$p(\text{objet}) = \frac{M_o}{M}$$

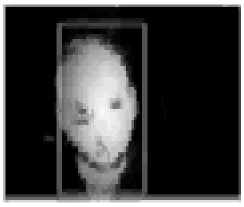
$$p(\vec{C}) = \frac{1}{M} h(\vec{C})$$

$$p(\vec{C} \mid \text{objet}) = \frac{1}{M_o} \; h_o(\vec{C})$$

thus

$$p(\text{objet} \mid \vec{C}) = p(\vec{C} \mid \text{objet}) \frac{p(\text{objet})}{p((\overset{\varnothing}{C}))} = \frac{1}{M_o} \; h_o(\vec{C}) \frac{\frac{M_o}{M}}{\frac{1}{M} h(\vec{C})} = \frac{h_o(\vec{c})}{h(\vec{c})}$$





3.3 Histograms of Receptive Field Values

This method can be generised to ANY vector of feautures. For example, the appearance of a neighborhood give by the receptive field vector.

$$\vec{V}(i,j;\sigma_i,\theta_i) = P(i,j) * (G_x, G_{xx}, G_{xy}, G_{yy}) \text{ at } \sigma_i \text{ and } \theta_i.$$

ATTENTION. The histogram must have sufficient samples M.

$$M \geq 10\ Q \geq 10\ N^{D.}$$

For the above exemple: D = 4.

Here is a table of numbers of cells in a histogram of D dimensions of N values.

N \ d	1	2	3	4	5	6
2	2^1	2^2	2^3	2^4	2^5	2^6
4	2^2	2^4	2^6	2^8	2^{10} =1 Kilo	2^{12} =2 K
8	2^3	2^6	2^9	2^{12}	2^{15}	2^{18}
16	2^4	2^8	2^{12}	2^{16}	2^{20} = 1 Meg	2^{24} = 4
32	2^5	2^{10} =1 Kilo	2^{15}	2^{20} = 1 Meg	2^{25}	2^{30} = 1
64	2^6	2^{12}	2^{18}	2^{24}	2^{30} = 1 Gig	2^{36}
128	2^7	2^{14}	2^{21} = 2 Meg	2^{28}	2^{35}	2^{42} =2 T
256	2^8	2^{16}	2^{24}	2^{32} = 2 Gig	2^{40} = 1 Tera	2^{48}

Consider the chromatic receptive fields normalized in scale and orientation σ_i and θ_i.

$$\vec{G}_{\sigma,\theta} = (G_x^L,\ G_x^{C_1},\ G_x^{C_2},\ G_{xx}^L,\ G_{xy}^L,\ G_{xx}^{C_1},\ G_{xx}^{C_2})$$

D= 7.

$$p(\text{objet}(i,j)\ |\ \vec{V}(i,j)\ ) = \frac{p(\vec{V}(i,j) \mid object(i,j)}{p(object(i,j))} p(\vec{V}(i,j) \approx \frac{h_o(\vec{V}(i,j))}{h(\vec{V}(i,j))}$$