High Level Vision

D.A. Forsyth
Computer Science Division,
U.C. Berkeley,
Berkeley,
CA 94720

Overview

- Topics
  - Recognition
  - Segmentation
  - Relational reasoning
  - Knowledge building
  - Generalization
- Sources of information
  - Images
    - single
    - multiple - depth, 3D structure
  - Video
    - tracking
  - Annotations
- Handoit contains more bibliographic info, suggested reading, etc

Applications - watching people

- Who is this?
  - face recognition at airports; deployed, but doesn't actually work
- Does the same person appear in many different places?
- Is someone behaving strangely?
  - where they shouldn't be (fairly easy)
  - wearing bulky explosives (seems to be hard)
  - doing something unusual (spectacularly hard?)
- What's happening?
  - What are the good bits of this surveillance video?

Applications - fishing in big datasets

- Iconic matching
  - child abuse prosecution
  - managing copyright (BayTSP)
- Clustering
  - Browsing for:
    - web presence for museums (Barnard et al, 01)
    - home picture, video collections
    - selling pictures
- Searching
  - scanned writing (Manmatha, 02)
- Building world knowledge
  - a face gazetteer (Miller et al, 04)

Model-based vision

- Problems
  - detection; localization; kinematics; counting
- Matching
  - Is this a pattern of a fixed class?
    - face detection
  - To what class does this pattern belong?
    - finding faces, animals, motorcycles, etc.
  - Primary issues:
    - local image representation
    - spatial representation
    - efficiency

Segmentation

- Problem
  - What components of the image likely belong together, and together form an object?
  - Can be thought of as like recognition of an unknown object
- Methods
  - clustering by
    - K means
    - EM
    - Graph theoretic methods

Relational reasoning

- Currently
  - Objects are composed of parts
    - Find the parts
    - Are the relations right?
- Perhaps
  - How are objects distributed in space?
  - Which objects are made of the same stuff?

Knowledge building

- Shop around mixed collections to obtain world knowledge
- Exploit the complementary nature of pictures, annotations
- e.g.
  - building object models
  - building a face dictionary
  - predicting who's in the picture
  - the cherimoya problem

Generalization

- Map knowledge across kinds of object
  - "This <animal> will butt or kick, but won't bite"
  - "This <animal> can bite, and is about to pounce"
- Requires
  - identifying "kind" (significant component is visual)
  - knowing what can be mapped, and where (mysterious)

The tetrad of vision

- Detection
  - what pictures contain a giraffe?
- Localization
  - where should I shoot to hit a giraffe?
- Kinematics
  - what is the giraffe's configuration?
- Counting
  - how many giraffes are there?

Detection

- Experimental protocol
  - apply detector to images known to contain/lack object, count
- Relatively easy to get performance figures
  - one doesn't need to check the giraffe has been put in the right place
  - but they may be meaningless or unreliable
  - in many test sets, objects and backgrounds are strongly correlated
- One should compare performance to baseline
  - e.g. SVM's on colour histograms; etc.
- Published performance figures are suspect
  - detection rates are implausibly high
  - datasets seldom baselined

Localization

- Experimental protocol unclear
  - how does one score partially correct localization?
  - errors are meaningful only wrt spatial model
- Experiments tricky on a respectable scale
  - but one or two images used to be common
- More difficult criterion to do well at than detection
  - can detect without localizing (detection marginalizes out configuration)
- Few published performance figures

Kinematics

- Experimental protocol thoroughly unclear
  - what is a partial success?
  - what does one count?
  - how?
- Not much known except for human tracking cases

Counting

- Experimental protocol easy in principle
- Obvioiusly, very difficult to do without localization
  - appears to be difficult even with models that can localize
- No current system can count anything significant satisfactorily

Basic template matching

- Core algorithm
  - Search image windows, present to a classifier, is this an x
- Issues
  - scale - search scales
  - lighting - correct for lighting
  - rotation - estimate rotation
  - variation in appearance, background - get a smarter classifier (?)
- Tremendous success in face finding

Rowley-Baluja-Kanade face finder (1)

Figure from "Rotation invariant neural-network based face detection,"
H.A. Rowley, S. Baluja and T. Kanade, Proc. Computer Vision and Pattern Recognition,
1998, c 1998, IEEE as shown in Forsyth and Ponce, p589

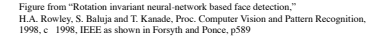


Figure from "Rotation invariant neural-network based face detection,"
H.A. Rowley, S. Baluja and T. Kanade, Proc. Computer Vision and Pattern Recognition,
1998, c 1998, IEEE as shown in Forsyth and Ponce, p589

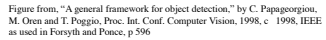


Figure from, "A general framework for object detection," by C. Papageorgiou,
M. Oren and T. Poggio, Proc. Int. Conf. Computer Vision, 1998, c 1998, IEEE
as used in Forsyth and Ponce, p 596

Difficulties

- Variation in appearance makes difficulties for the classifier
  - Sources
    - pose; aspect; lighting; within-class variation; kinematic degrees of freedom; segmentation
  - templates work best when "implicit" segmentation is easy
- We've ignored
  - Differences between classifiers
    - now an enormous literature of different face finders, see, e.g.
  - Methods to build very fast or very efficient classifiers
    - fairly large literature on this, see, e.g.,
  - Feature selection
    - not much organized literature yet, but see, e.g.,

More complex template matching

- Encode an object as a set of patches
  - centered on interest points
  - match by
    - voting
    - spatially censored voting
    - inference on a spatial model

We'll see these cases when we talk about matching on relations.

Pinhole camera (F+P, p31)

Perspective camera (F+P, p33)

Orthographic camera (F+P, p33)

View variation for a plane patch

- Plane patches look different in different views
  - Perspective views induce a homography
  - Scaled orthographic views induce an affine transformation

Interest points and local descriptions

- Find localizable points in the image
  - e.g. corners
- Build
  - a local coordinate frame
    - Euclidean+scale
    - Affine
  - a representation of the image within that coordinate frame

Belongie/Malik shape contexts

Lowe's keypoints and SIFT features

Mikolaczyk/Schmid coordinate frames

Matching objects by voting on keypoints

Figure from "Local grayvalue invariants for image retrieval," by
C. Schmid and R. Mohr, IEEE Trans. Pattern Analysis and Machine Intelligence, 1997
c 1997, IEEE as used in Forsyth + Ponce, p 609

Views of 3D objects

- Important, somewhat interrelated phenomena
  - Visibility
    - different subsets of an object can be seen from different viewing directions
  - Aspect
    - Objects look different when seen from different directions
    - Crucial fact:
      - outline points derived from vertices, sharp edges don't move on the surface
      - outline points derived from smooth points do move on the surface

Contour generator and outline for perspective and orthography, (F+P, p485,499)

Viewpoint Consistency

- General principle:
  - all features are viewed in the same camera
- Most common form:
  - hypothesize a model, some model-image feature correspondences
  - calibrate the camera using correspondences
  - project other features into the image using calibrated camera
  - confirm/reject hypothesis by testing neighbourhood of projected features
- Variants:
  - camera representation, parameters recovered, features employed, testing strategy
- Key notion:
  - frame-bearing feature group

Figure from "Object recognition using alignment," D.P. Huttenlocher and S. Ullman, Proc. Int. Conf. Computer Vision, 1986 as used in Forsyth and Ponce, p459

View consistency and voting

- Variant: vote on camera parameters
  - for all models, model-image fbfg correspondences
    - compute camera parameters using correspondences
    - add vote to relevant bucket
  - the camera parameters are given by the bucket with the most votes
  - the object can be read off the bucket
- Generally, not a good thing --- noise

Figure from "The evolution andtesting of a model-based object recognition system", J.L. Mundy and A. Heller, Proc. Int.Conf. Computer Vision, 1990 c 1990 IEEE, Forsyth and Ponce p465

Hypothesis verification

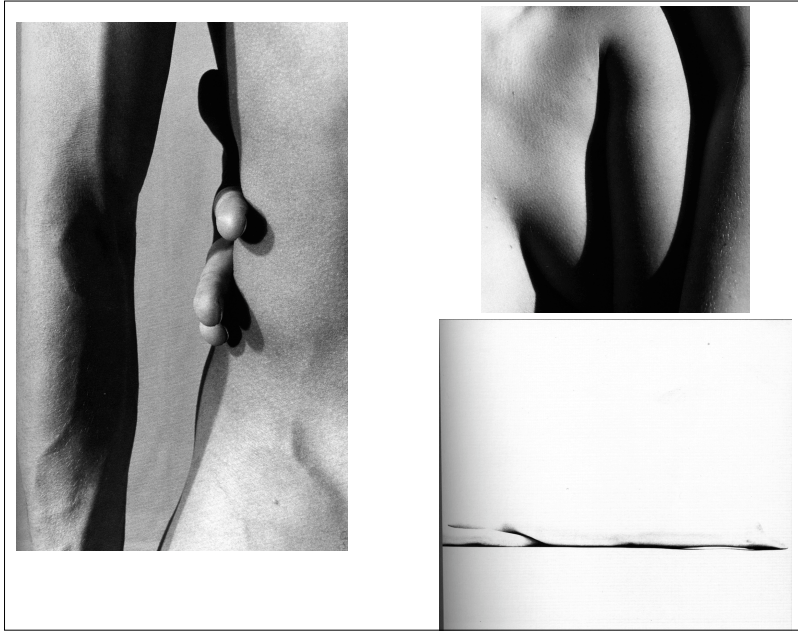
- Backproject and score
- Current:
  - score edge points
  - score oriented edge points
- Desirable:
  - score "similarity in appearance"
  - given "context of images"
- A natural, but unexplored, domain for learning methods
  - discussing verification is currently somewhat unfashionable
  - pretty much every recognition system has (and will have?) a verification step.

Difficulties in verification

from "Efficient model library access by projectively invariant indexing functions," by C.A. Rothwell et al., Proc. Computer Vision and Pattern Recognition, 1992, c 1992, IEEE, given on p475 of Forsyth+Pon ce

Viewpoint consistency can be made to work for curved objects

Figure from "On Recognising and positioning curved 3d objects from image contours," D.J. Kriegman and J. Ponce, IEEE Trans. Pattern Analysis and Machine Intelligence, 1990, c IEEE 1990 From Forsyth and Ponce, p482



Segmentation and grouping

- Motivation: not all image information is evidence
- Notion: an informative representation from image/video
- General ideas:
  - Segmentation: decompose image into informative domains
  - Grouping: cluster together tokens that "belong together"
  - Tokens: whatever we might need to group (points, patches, etc.)
  - Top-down: belong together because they lie on the same model
  - Bottom-up: belong together because they are locally coherent.

Basic ideas of grouping in humans

- Figure-ground segregation
  - allocate some elements to figure, some to ground
  - impoverished theory
- Gestalt properties
  - elements in a collection of elements can have properties that result from relationship (e.g. Muller Lyer effect)
    - gestaltqualitat
  - A series of factors affect whether elements should be grouped together

Gestalt factors

Segmentation as clustering

- Represent each image pixel with a vector of attributes
  - e.g. grey level, color, position, smoothed energy in filter responses, etc.
- Cluster these vectors
- Backproject to image

- Natural clustering strategies
  - k-means
  - EM on a Gaussian mixture model
  - spectral clustering of various forms

K-means

- Fix a number of clusters
- Iterate
  - fix cluster centers, allocate points to closest center
  - fix allocations, compute best cluster centers
- Minimizes

EM on a Gaussian mixture model

- Hardly bears detailed description
  - significant feature -- smooth allocation to clusters
  - wierd nasty fact -- often worse than EM

Spectral clustering methods

- Pixels are nodes in a weighted graph
- Edges are weighted with affinity
- Cut this graph

- Affinity measures combine
  - intensity
  - distance
  - colour
  - texture
  - motion

Simplest spectral clustering

Normalised cuts

Segmentation seeking primitives

- Segments of interest may have a particular, simple form
- E.g. people and animals=cylinders=rectangles
- Build task specific segmenter
  - similar in spirit to template matcher

Body segments by temporal coherence

Body segments from constraints

We didn't discuss

- Evaluation
  - compare with human segmentations
  - use for some practical application
- Efficiency
  - numerous tricks available for computing eigenvectors fast
- Fitting
  - assembling tokens to form geometric primitives
- Pros and cons of various representations

Parts and wholes

- Generally
  - objects are made up of parts
  - detect parts; if they're in the right configuration, object is there
- Image-3D relations
  - it would be nice if objects were made of parts whose outlines were strongly constrained
- Part-part relations
  - kinematics of parts in the image is constrained by kinematic constraints in 3D

Image-part relations

- Cylinders
  - view is usually orthographic
  - outline consists of two parallel lines
- Generalized cylinders
  - controversial, somewhat fluffy idea
  - many objects are "swept", resulting in "swept" outlines
- Straight homogenous generalized cylinders
  - cylinder with non-circular cross-section
  - outline consists of multiple parallel lines
- Various fluffier cases
  - geons, etc.

Forsyth+Ponce, p647

Forsyth+Ponce, p648

Forsyth+Ponce, p654

Figure from "Segmentation and description based on perceptual organisation," R. Mohan and R. Nevatia, Proc. Computer Vision and Pattern Recognition, 1989, c 1989, IEEE as used in Forsyth and Ponce, p657,658

Inferring 3D kinematics from body segments

- Body segments are cylinders of (roughly) known length
- Views are (essentially) scaled orthography
- Hence, from the image length one gets cos(slant)
- This allows 3D reconstruction

Text

Figure from "Reconstruction of Articulated Objects from Point Correspondences in a Single Uncalibrated Image," C.J. Taylor, Proc. Computer Vision and Pattern Recognition, 2000, c 2000, IEEE; as used in Forsyth+Ponce, p 662

Figure from "Reconstruction of Articulated Objects from Point Correspondences in a Single Uncalibrated Image," C.J. Taylor, Proc. Computer Vision and Pattern Recognition, 2000, c 2000, IEEE as used in Forsyth+Ponce, p663

Part-part relations

- Co-occurrence
  - e.g. voting
- Spatial relations
  - constrained voting
  - kinematic grouping
  - constellation models

Simplest co-occurrence

- Describe local interest points, as before
- Cluster interest point descriptors
- Each descriptor votes for every object that contains one such
- Object with the most votes, wins

Figure from "Local grayvalue invariants for image retrieval," by C. Schmid and R. Mohr, IEEE Trans. Pattern Analysis and Machine Intelligence, 1997 c 1997, IEEE as used in Forsyth + Ponce, p 609

Co-occurrence with geometric censor

Figure from "Local grayvalue invariants for image retrieval," by C. Schmid and R. Mohr, IEEE Trans. Pattern Analysis and Machine Intelligence, 1997 c 1997, IEEE as used in Forsyth + Ponce, p 612

Figure from "Local grayvalue invariants for image retrieval," by C. Schmid and R. Mohr, IEEE Trans. Pattern Analysis and Machine Intelligence, 1997 c 1997, IEEE as used in Forsyth + Ponce, p 613

Co-occurrence using object statistics

- Freeman et al

Kinematic grouping

- Assemble a set of features to present to a classifier
  - which tests
    - appearance
    - configuration
    - whatever
- Classifier could be
  - handwritten rules (e.g. Fleck-Forsyth-Bregler 96)
  - learned classifier (e.g. Ioffe-Forsyth 99)
  - likelihood (e.g. Felzenszwalb-Huttenlocher 00)
  - likelihood ratio test (e.g. Leung-Burl-Perona 95; Fergus-Perona-Zisserman 03)

Kinematic grouping

- Three questions
  - given a group, what should be added?
  - can a given group be pruned without adding anything further?
  - can a given group be accepted without adding anything further?
- All three are tied up with the form of the final classifier
- All three resonate with “classical” issues
  - FBFG; camera consistency; verification

Pictorial structures

- For models with the right form, one can test “everything”
  - model is a set of segments linked into a tree structure
  - putative image segments are quantized
  - => dynamic programming to search all matches
  - What to add next? (DP deals with this)
  - Pruning? (Irrelevant)
  - Can one stop? (Use a mixture of tree models, with missing segments marginalized out)
  - Known segment colour - Felzenszwalb-Huttenlocher 00
  - Learned models of colour, layout, texture - Ramanan Forsyth 03, 04

Figure from "Efficient Matching of Pictorial Structures,"
P. Felzenszwalb and D.P. Huttenlocher, Proc. Computer Vision and Pattern Recognition
2000, c 2000, IEEE as used in Forsyth+Ponce, p640

What should be added?

- For models with the right

Can we stop assembly?

- Equivalent to:
  - is it worth verifying this hypothesis?
  - is this assembly sufficient to assert object is present?
- Derived from:

Figure from "Body Plans," by D.A. Forsyth and M.M. Fleck, Proc. Computer Vision and Pattern Recognition,
1997, c 1997, IEEE as used in Forsyth+Ponce, p 620

Figure from Forsyth+Ponce, p 619

- Pruning strategy:
  - prune assemblies which cannot pass the classifier, whatever is attached
  - equivalent to projecting the decision boundary
  - can be repeated in stages

kinematic grouping ala lazebnick

Constellation models

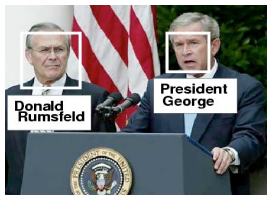
Knowledge building

- Use multiple components of a collection to build models
  - e.g. images and associated captions
- Use multiple collections to build models
  - opportunistically
  - link partial models via matching
  - e.g. spatial models on video, texture models from named collection, names from named collection
- No overarching theory yet
  - but seems like a quite useful idea

News dataset

- Approx 5e5 news images, with captions
  - Easily collected by script from Yahoo over the last 18 months or so
- Mainly people
  - politicians, actors, sportsplayers
  - long, long tails distribution
- Face pictures captured “in the wild”
- Correspondence problem
  - some images have many (resp. few) faces, few (resp. many) names (cf. Srihari 95)

President George W. Bush makes a statement in the Rose Garden while Secretary of Defense Donald Rumsfeld looks on, July 23, 2003. Rumsfeld said the United States would release graphic photographs of the dead sons of Saddam Hussein to prove they were killed by American troops. Photo by Larry Downing/Reuters

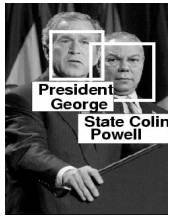


Process

- Extract proper names — Scale
  - rather crudely, at present
- Detect faces — 44773 big face responses
  - with Cordelia Schmid's face detector, (Vogelhuber Schmid 00)
- Rectify faces — 34623 properly rectified
  - by finding eye, nose, mouth patches, affine transformation
- Kernel PCA rectified faces
- Estimate linear discriminants
- Now have (face vector; name_1,...., name_k) — 27742 for k<=4

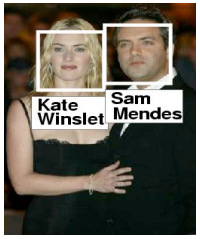
Building a face dictionary

- Compute linear discriminants
  - using single name, single face data items
  - we now have a set of clusters
-
- Now break correspondence with modified k-means
  - assign face to cluster with closest center,
    - chosen from associated names
  - recompute centers, iterate
  - using distance in LD space
-
- Now recompute discriminants, recluster with modified k-means

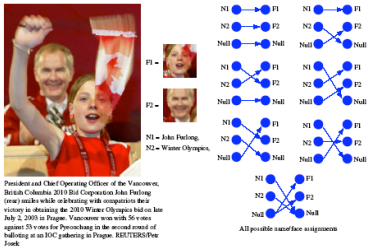


US President George W. Bush (L) makes remarks while Secretary of State Colin Powell (R) listens before signing the US Leadership Against HIV /AIDS , Tuberculosis and Malaria Act of 2003 at the Department of State in Washington, DC. The five-year plan is designed to help prevent and treat AIDS, especially in more than a dozen African and Caribbean nations(AFP/ Luke Frazza)

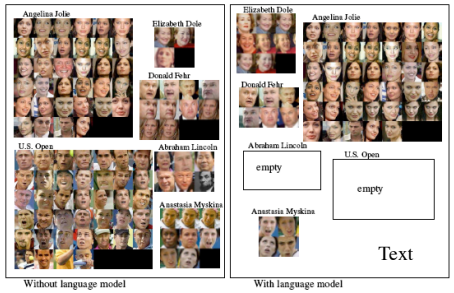
German supermodel Claudia Schiffer gave birth to a baby boy by Caesarian section January 30, 2003, her spokeswoman said. The baby is the first child for both Schiffer, 32, and her husband, British film producer Matthew Vaughn, who was at her side for the birth. Schiffer is seen on the German television show 'Bet It...?!' ('Wetten Dass...?!') in Braunschweig, on January 26, 2002. (Alexandra Winkler/Reuters)



British director Sam Mendes and his partner actress Kate Winslet arrive at the London premiere of 'The Road to Perdition', September 18, 2002. The films stars Tom Hanks as a Chicago hit man who has a separate family life and co-stars Paul Newman and Jude Law. REUTERS/Dan Chung



from "Who's in the picture," Berg, Berg, Edwards and Forsyth, in review



from "Who's in the picture," Berg, Berg, Edwards and Forsyth, in review

IN Pete Sampras IN of the U.S. celebrates his victory over Denmark's **OUT Kristian Pl**
OUT at the **OUT U.S. Open OUT** at Flushing Meadows August 30, 2002. Sampras won
match 6-3 7- 5 6-4. REUTERS/Kevin Lamarque

Germany's **IN Chancellor Gerhard Schroeder IN**, left, in discussion with France's **IN Pre**
dent Jacques Chirac IN on the second day of the EU summit at the European Council headqu
ters in Brussels, Friday Oct. 25, 2002. EU leaders are to close a deal Friday on finalizing en
talks with 10 candidate countries after a surprise breakthrough agreement on Thursday betwe
France and Germany regarding farm spending.(AP Photo/European Commission/HO)

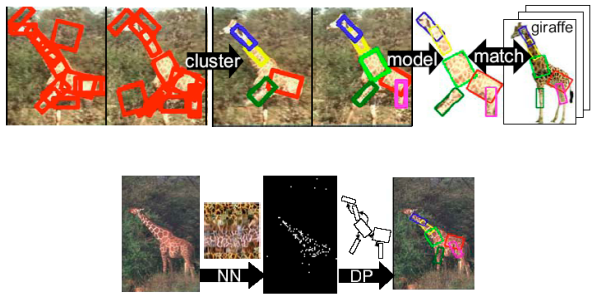
'The Right Stuff' cast members **IN Pamela Reed IN**, (L) poses with fellow cast member
Veronica Cartwright IN at the 20th anniversary of the film in Hollywood, June 9, 2003. T
women played wives of astronauts in the film about early United States test pilots and the spa
program. The film directed by **OUT Philip Kaufman OUT**, is celebrating its 20th anniversa
and is being released on DVD. REUTERS/Fred Prouser

Kraft Foods Inc., the largest U.S. food company, on July 1, 2003 said it would take steps, l
capping portion sizes and providing more nutrition information, as it and other companies fa
growing concern and even lawsuits due to rising obesity rates. In May of this year, San Fra
cisco attorney **OUT Stephen Joseph OUT**, shown above, sought to ban Oreo cookies in Ca
fornia – a suit that was withdrawn less than two weeks later. Photo by Tim Wimborne/Reut
REUTERS/Tim Wimborne

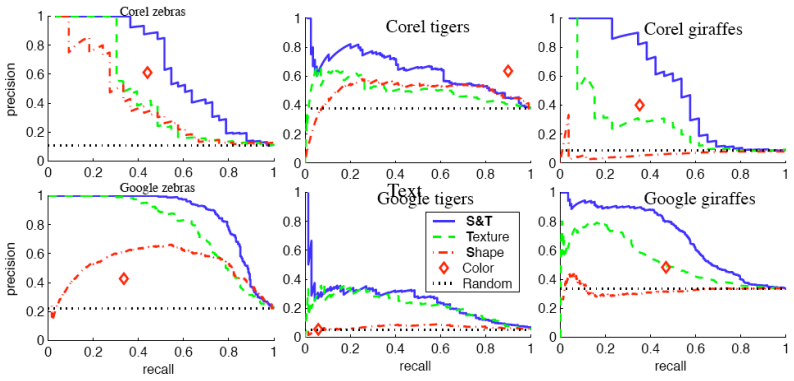
How well does it work?

- Draw a cluster from the list, and an image from that cluster
  - frequency that that image is of someone else
- How many bits are required to fix result?

#Images	#Clusters	error rate
19355	2357	26%
7901	1510	11%
4545	765	5.2%
3920	725	7.5%
2417	328	6.6%



From "Combining models for object recognition," Ramanan and Forsyth, in review



From "Combining models for object recognition," Ramanan and Forsyth, in review