Machine Learning in Computer Vision

A Tutorial

Ajay Joshi, Anoop Cherian and Ravishankar Shivalingam
Dept. of Computer Science, UMN

Outline

- Introduction
- Supervised Learning
- Unsupervised Learning
- Semi-Supervised Learning
  - Constrained Clustering
  - Distance Metric Learning
  - Manifold Methods in Vision
  - Sparsity based Learning
  - Active Learning
- Success stories
- Conclusion

Computer Vision and Learning

"The typical way people from learning look at computer vision"

Vision and Learning

Vision specific constraints/assumptions

Application of Learning Algorithms

Why Machine Learning?

Most of the real world problems are:

a) NP-Hard (ex: scene matching).
b) Ill-defined (ex: 3D reconstruction from a single image).
c) The right answer is subjective (ex: segmentation).
d) Hard to model (ex: scene classification)

Machine Learning tries to use statistical reasoning to find approximate solutions for tackling the above difficulties.

What kind of Learning Algorithms?

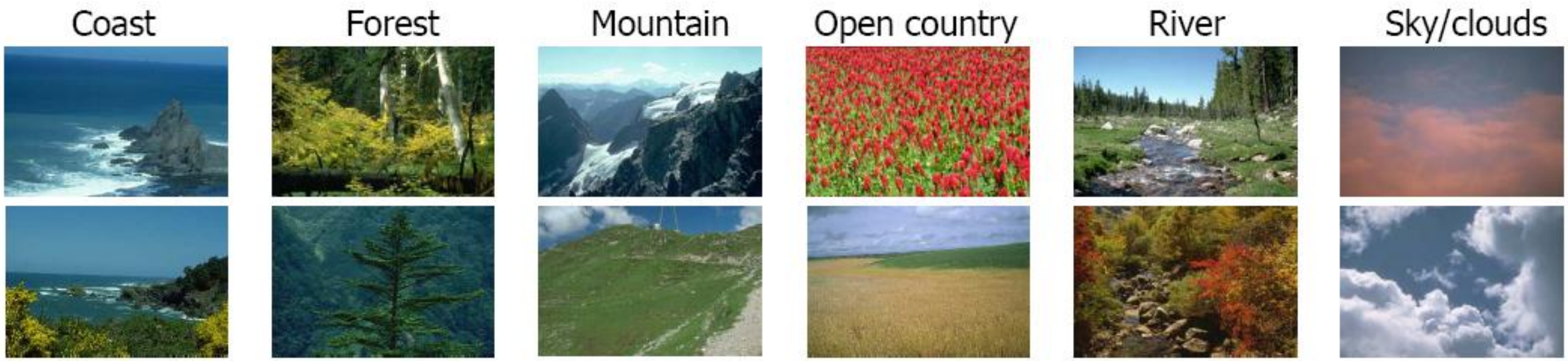
- Supervised Learning
  - Generative/Discriminative models
- Unsupervised Learning
  - K-Means/Dirichlet/Gaussian Processes
- Semi-Supervised Learning
  - The latest trend in ML and the focus of this tutorial.

Supervised Learning

- Uses training data with labels to learn a model of the data
- Later uses the learned model to predict test data.
- Traditional Supervised learning techniques:
  - Generative Methods
    - Naïve Bayes Classifier
    - Artificial Neural Networks
    - Principal Component Analysis followed by Classification, etc.
  - Discriminative methods
    - Support Vector Machines
    - Linear Discriminant Analysis, etc.

Example: Scene Classification

- Given a corpora of sample data of various scenes and their associated labels, classify the test data.

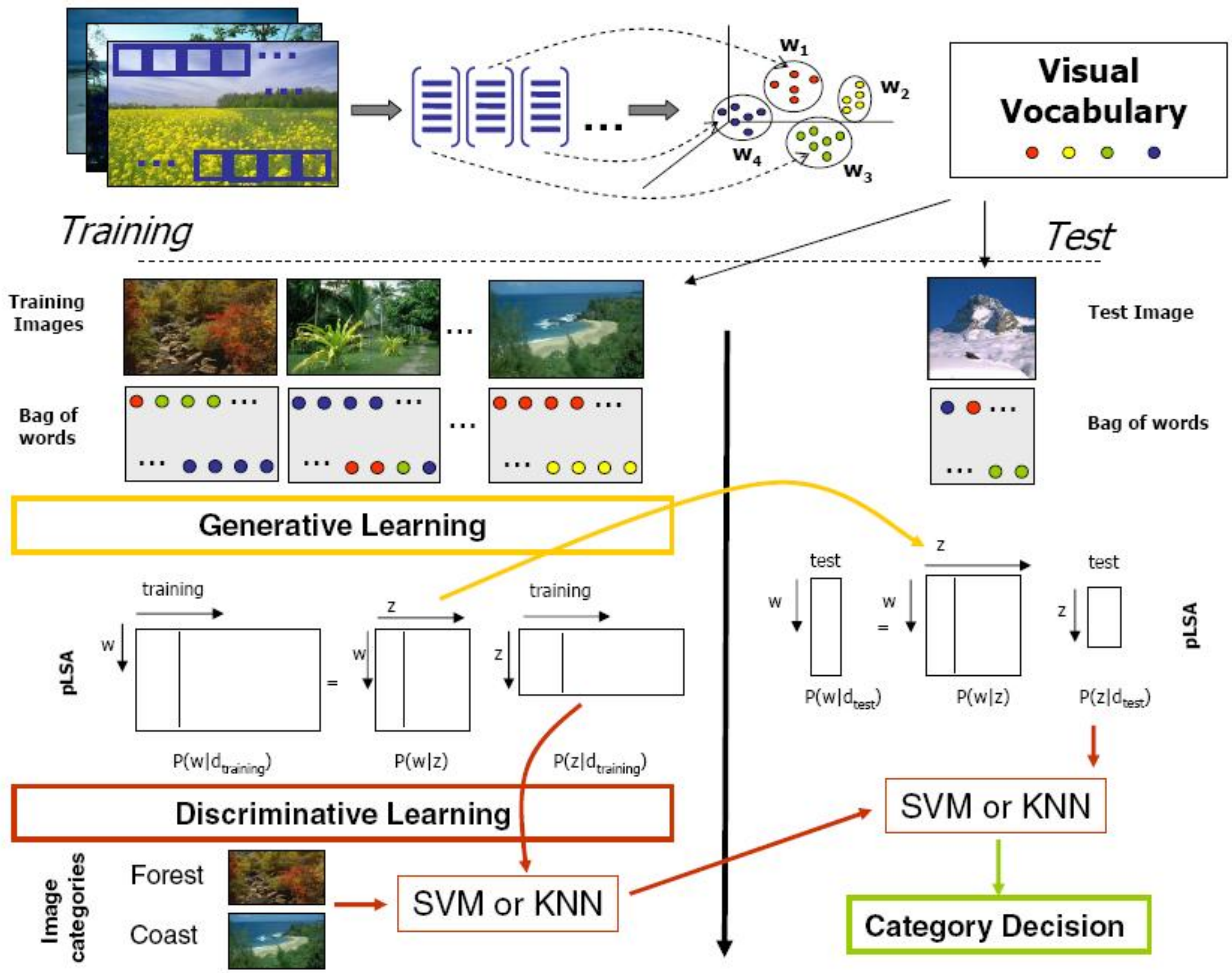


Training data with labels.

Scene Classification Continued…

- One way to do this:
  - Using a combination of Generative and Discriminative Supervised Learning models (Zissermann, PAMI'09).
  - Divide the training images into patches.
  - Extract features from the patches and form a dictionary using Probabilistic Latent Semantic Analysis.
    - Consider image as a document d, with a mixture of topics z and words d. Decide the possible number of topics pre-hand.
    - $P(w|d) = \sum_{z \epsilon Z} P(w|z)P(z|d)$
  - Use EM on the training data to find *P(w|z)* and *P(z|d).*
  - Train a discriminative classifier (SVM) on *P(z|d)* and classify test images.

Scene Classification Algorithm



Supervised Learning: Problems

- Unavailability of labeled data for training the classifier
  - Labeling data is boring
  - Experts might not be available (ex: medical imaging).

- Number of topic categories might not be available (as in the case of scene classification mentioned earlier) or might increase with more data.

- Solution: Unsupervised Learning.

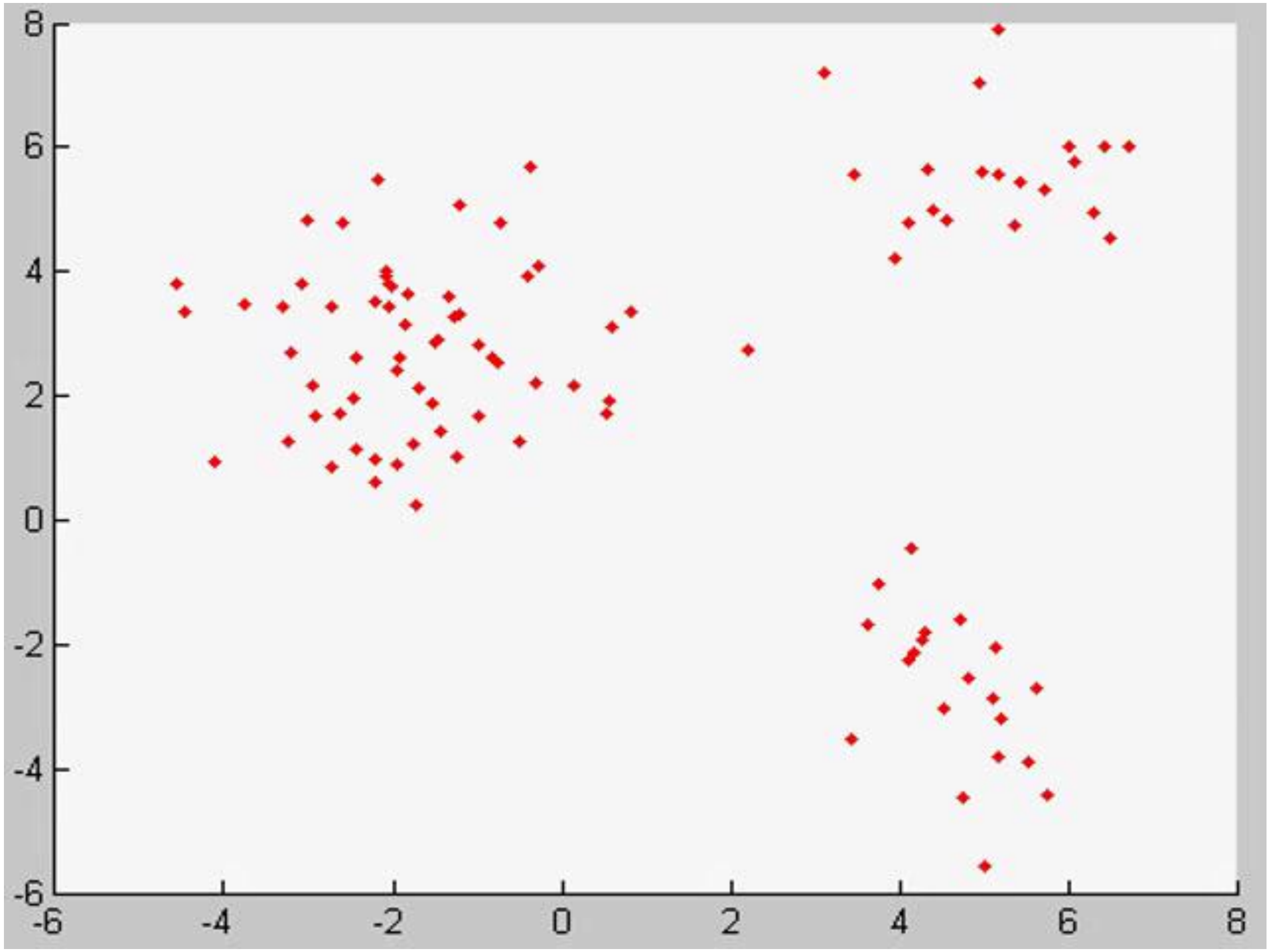
Unsupervised Learning

- Learner is provided only unlabeled data.
- No feedback is provided from the environment.
- Aim of the learner is to find patterns in data which is otherwise observed as unstructured noise.
- Commonly used UL techniques:
  - Dimensionality reduction (PCA, pLSA, ICA, etc).
  - Clustering (K-Means, Mixture models, etc.).

Non-Parametric clustering techniques

- In the previous Scene Classification example, what if we do not know the number of scene topics, z, available in the data?
- One possibility is to use Dirichlet Process Mixture Models (DPMM) for clustering.
  - Data is assumed to be samples from by an infinitely parameterized probability distribution.
  - Dirichlet Processes have the property that they can represent mixtures of infinite number of probability distributions.
  - Sample data from DPMM and try to fit the best clustering model that can explain the data.

Non-parametric model learning using Dirichlet Processes



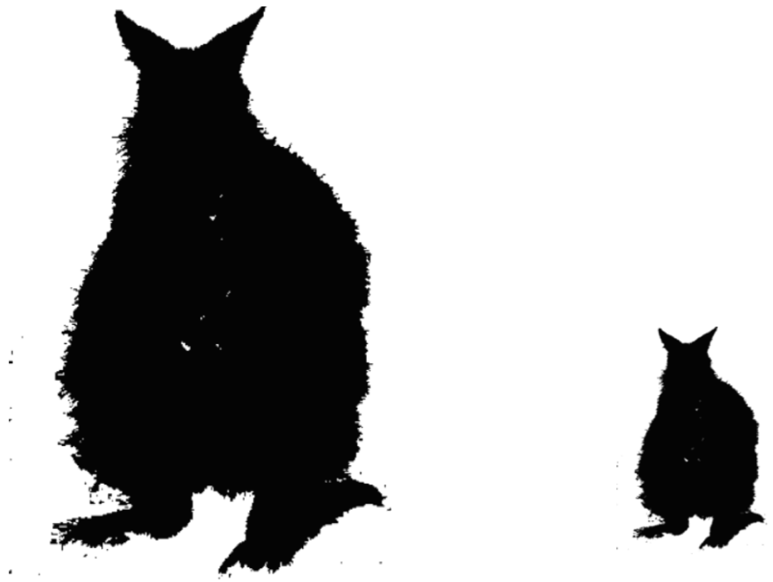
(Video)

Unsupervised Learning: Problems

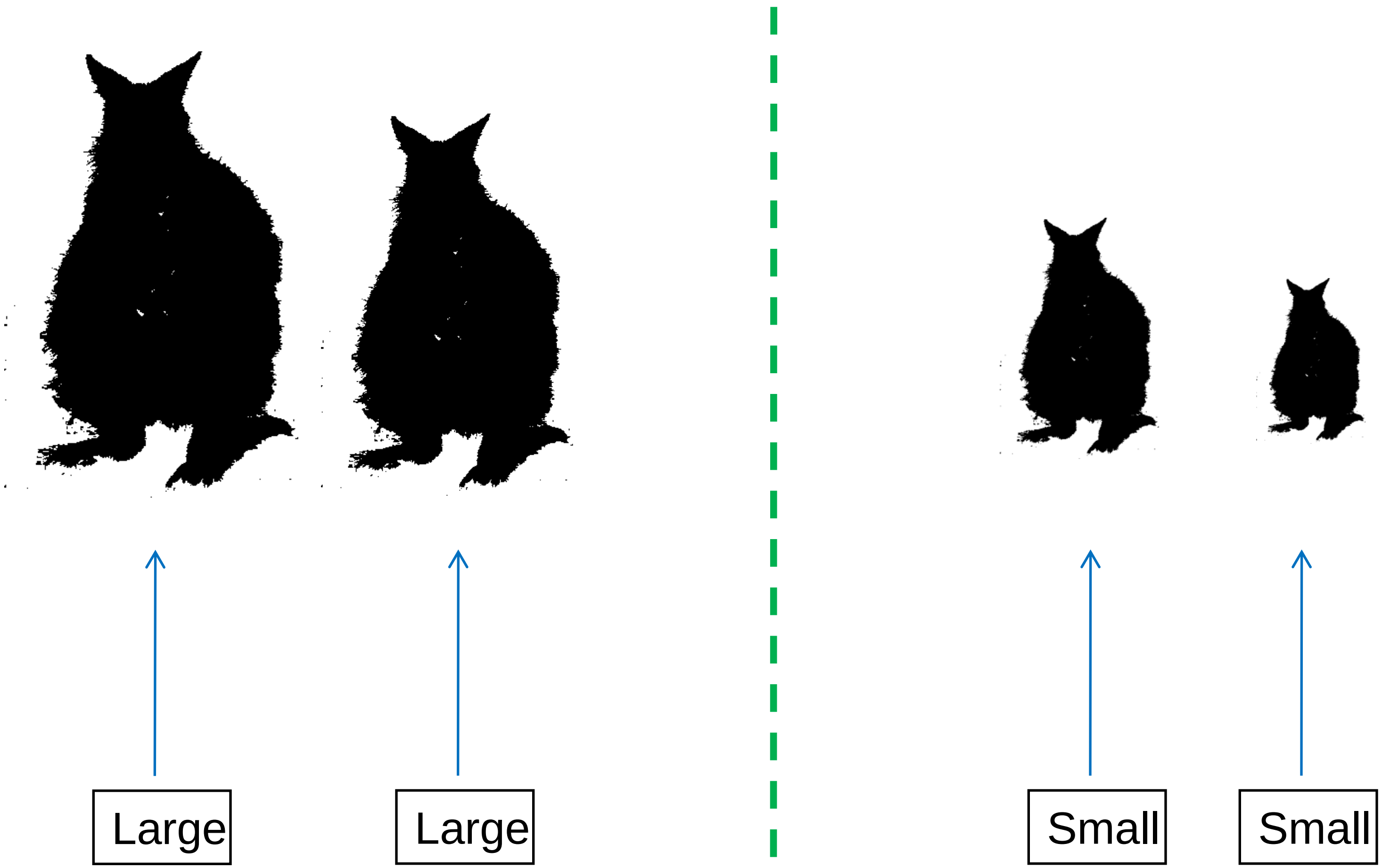
- Clusters generated by unsupervised learners might not adhere with real world clustering.
- Real world problems are often subjective. Ex: segmentation.
- Can a little bit of labeled data be used to guide an unsupervised learner?
- Can the learner incorporate user suggestions and feedback?
- Solution: Use Semi-Supervised Learning (SSL).

SSL: A motivating Example

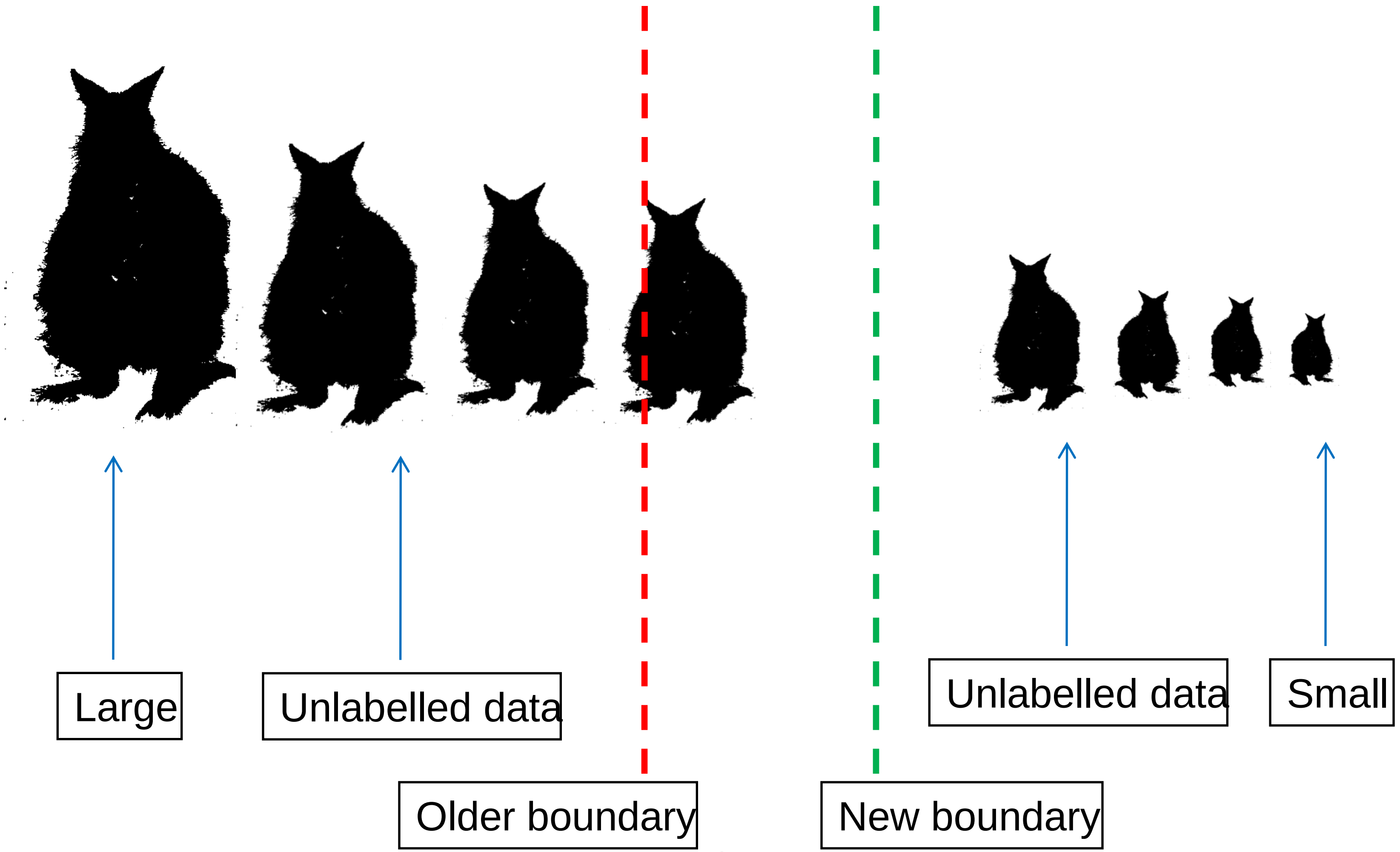
Classify animals into categories of large and small!



Supervised Learning Approach



Semi Supervised Learning Approach



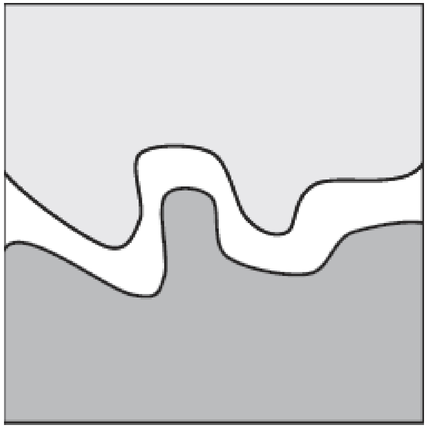
What is SSL?

- As the name suggests, it is in between Supervised and Unsupervised learning techniques w.r.t the amount of labelled and unlabelled data required for training.
- With the goal of reducing the amount of supervision required compared to supervised learning.
- At the same time improving the results of unsupervised clustering to the expectations of the user.

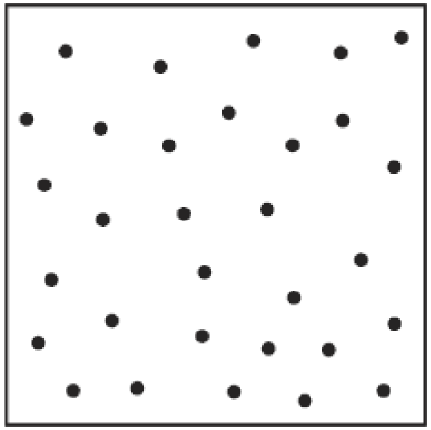
Assumptions made in SSL

- Smoothness assumption:
  - The objective function is locally smooth over subsets of the feature space as depicted by some property of the marginal density.
    - Helps in modeling the clusters and finding the marginal density using unlabelled data.
- Manifold assumption:
  - Objective function lies in a low dimensional manifold in the ambient space.
    - Helps against the curse of dimensionality.

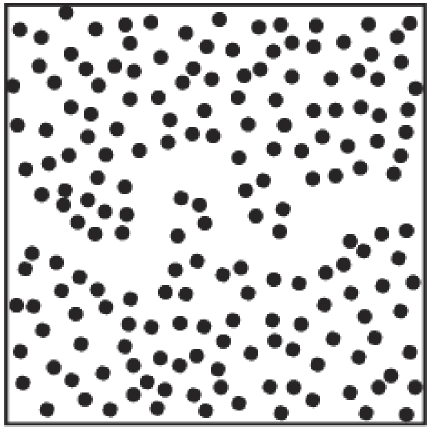
Learning from unlabelled data



Original decision boundary



When only labeled data is Given.



With unlabeled data along with labeled data

With lots of unlabeled data the decision boundary becomes apparent.

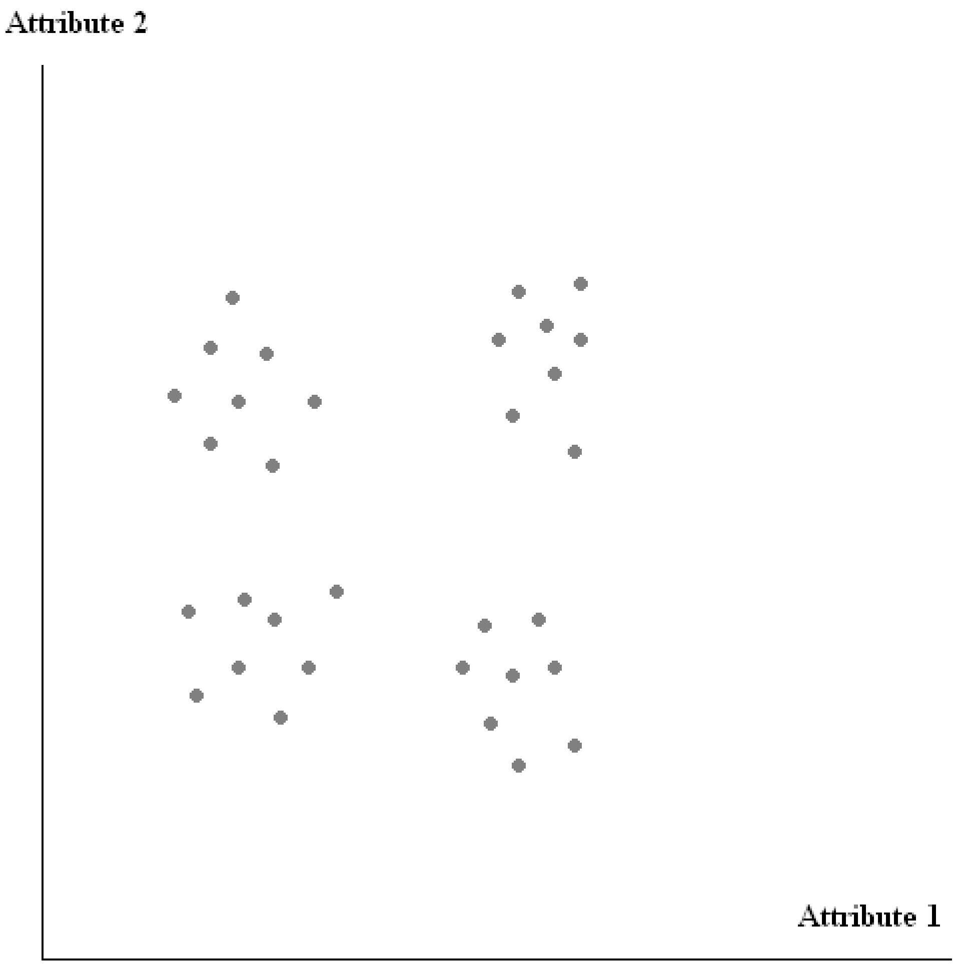
Overview of SSL techniques

- Constrained Clustering
- Distance Metric Learning
- Manifold based Learning
- Sparsity based Learning (Compressed Sensing).
- Active Learning

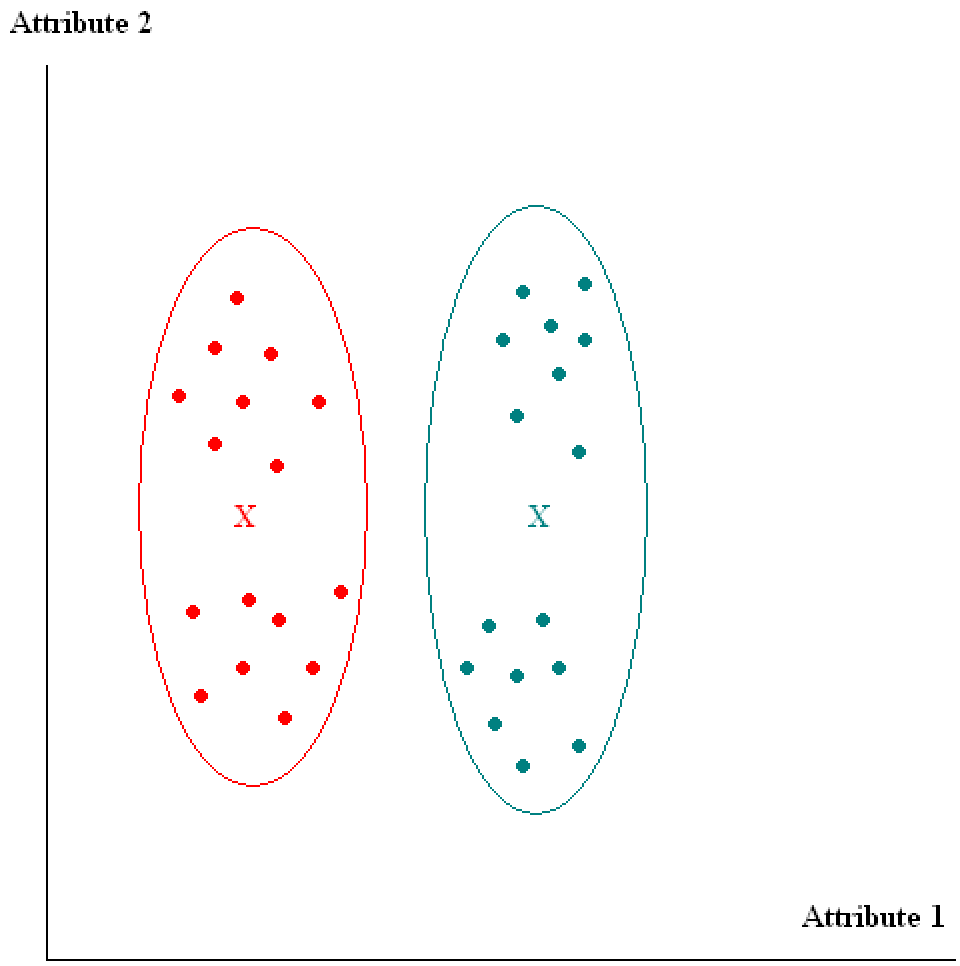
Constrained Clustering

- When we have any of the following:
  - Class labels for a subset of the data.
  - Domain knowledge about the clusters.
  - Information about the ‘similarity’ between objects.
  - User preferences.
- May be pairwise constraints or a labeled subset.
  - ***Must-link*** or ***cannot-link*** constraints.
  - Labels can always be converted to pairwise relations.
- Can be clustered by searching for partitionings that respect the constraints.
- Recently the trend is toward similarity-based approaches.

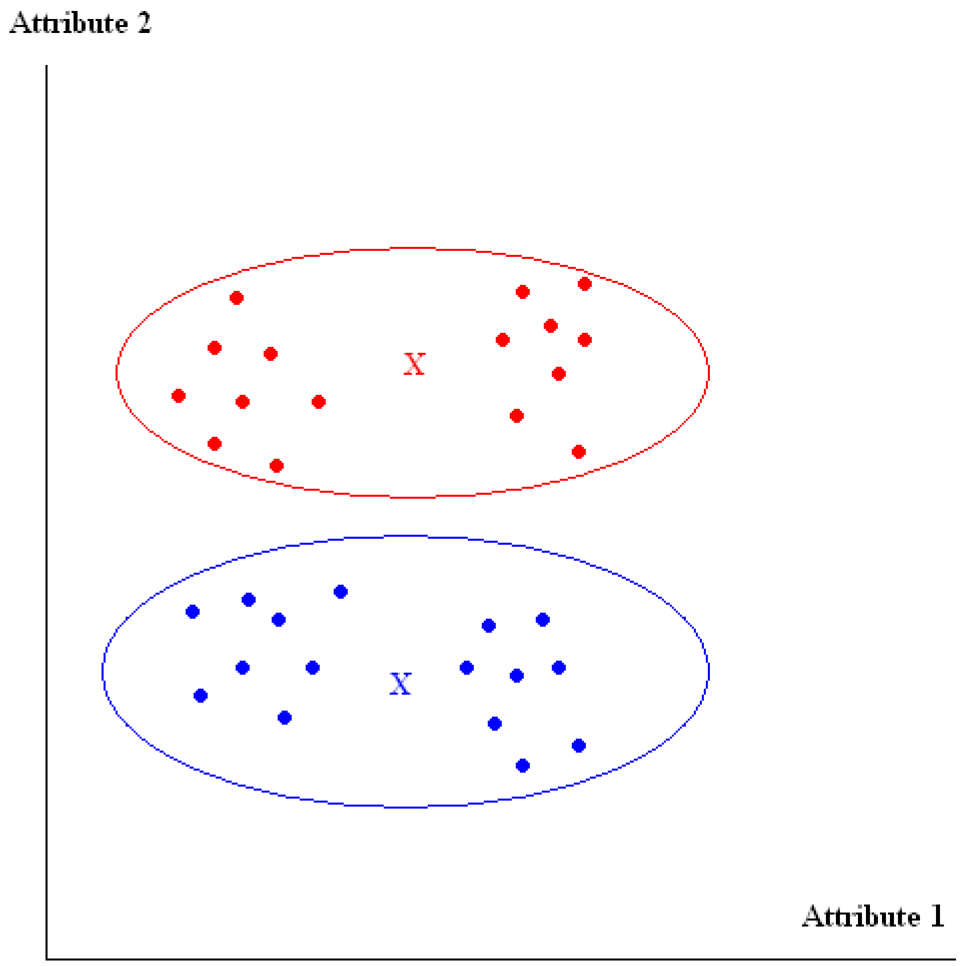
Sample Data Set



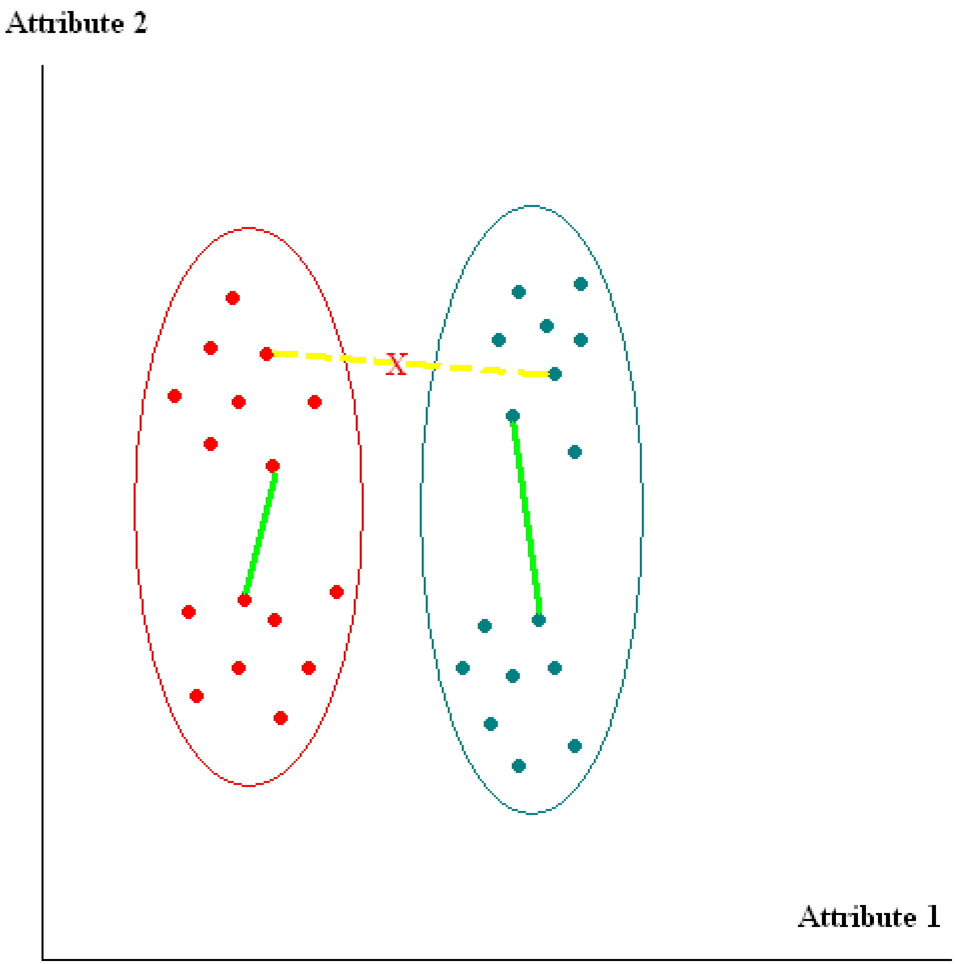
Partitioning A



Partitioning B



Constrained Clustering



Distance Metric Learning

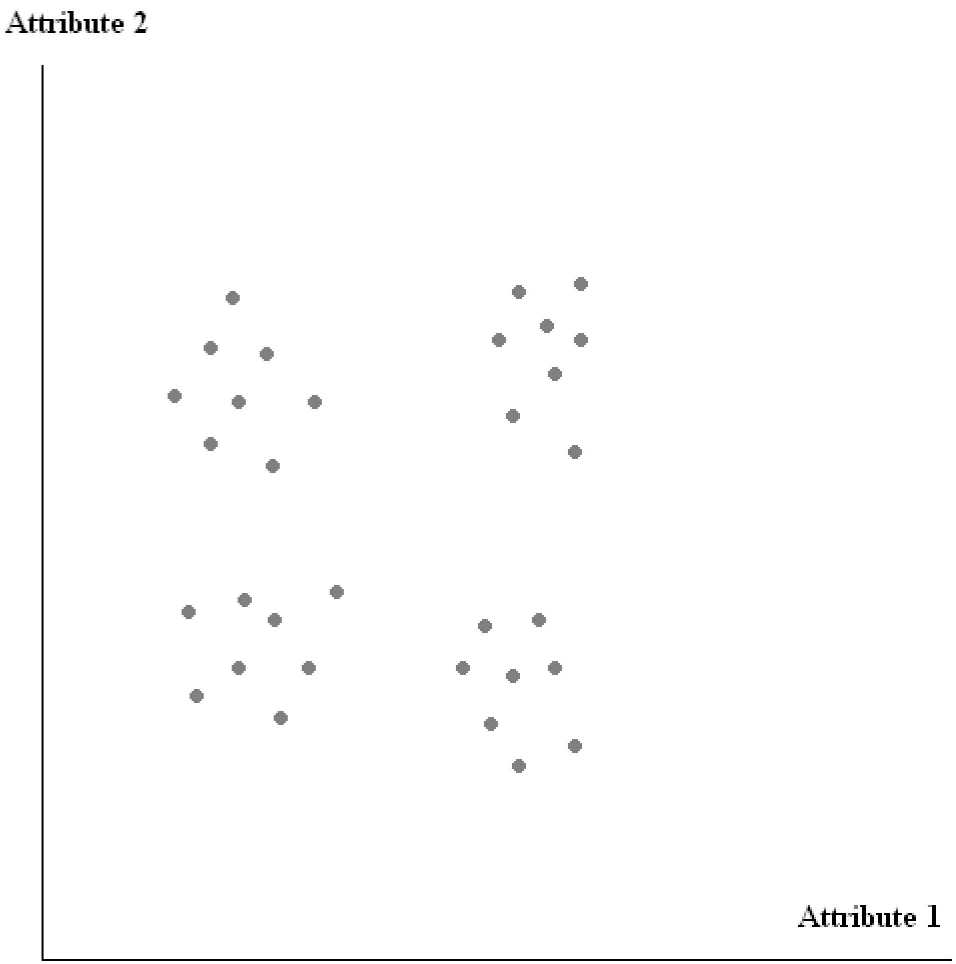
- Learning a 'true' similarity function, a distance metric that respects the constraints
- Given a set of pairwise constraints, i.e., must-link constraints $\mathcal{M}$ and cannot-link constraints $\mathcal{C}$
- Find a distance metric $\mathcal{D}$ that
  - Minimizes total distance between must-linked pairs

$$\sum_{(x,y)\in M} D(x,y)$$

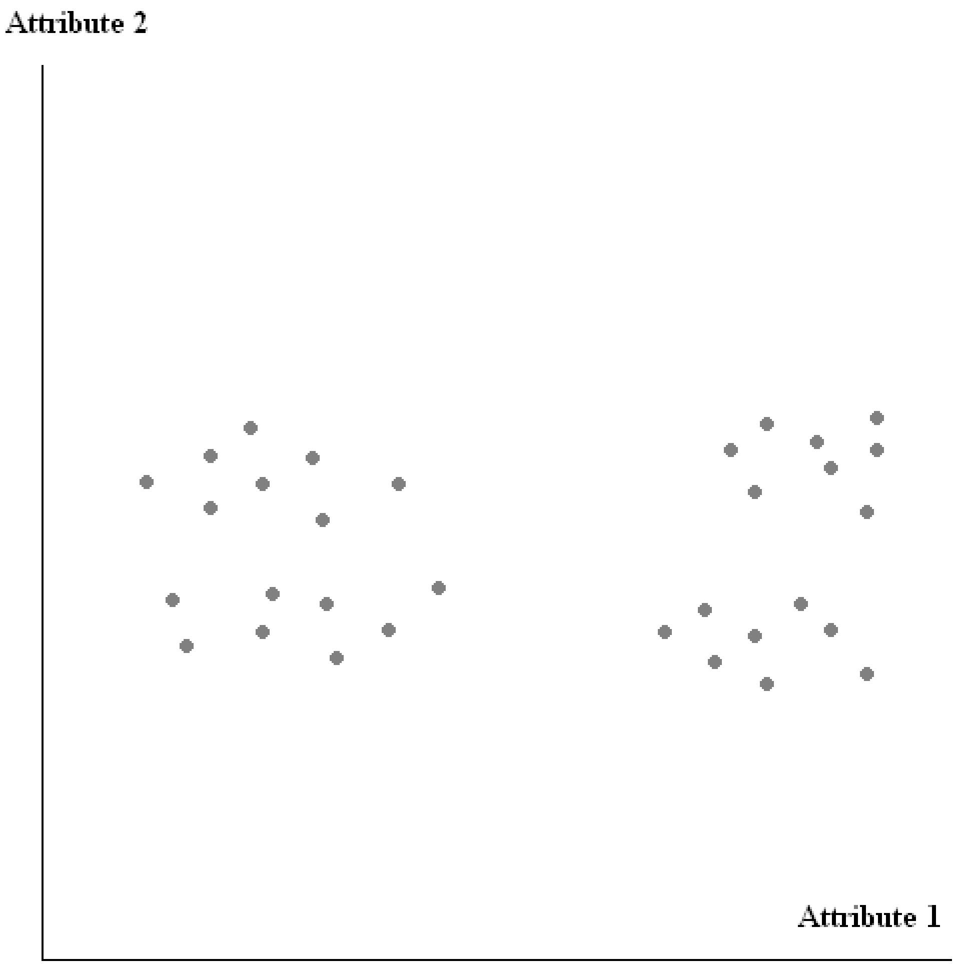
  - Maximizes total distance between cannot-linked pairs

$$\sum_{(x,y)\in C} D(x,y)$$

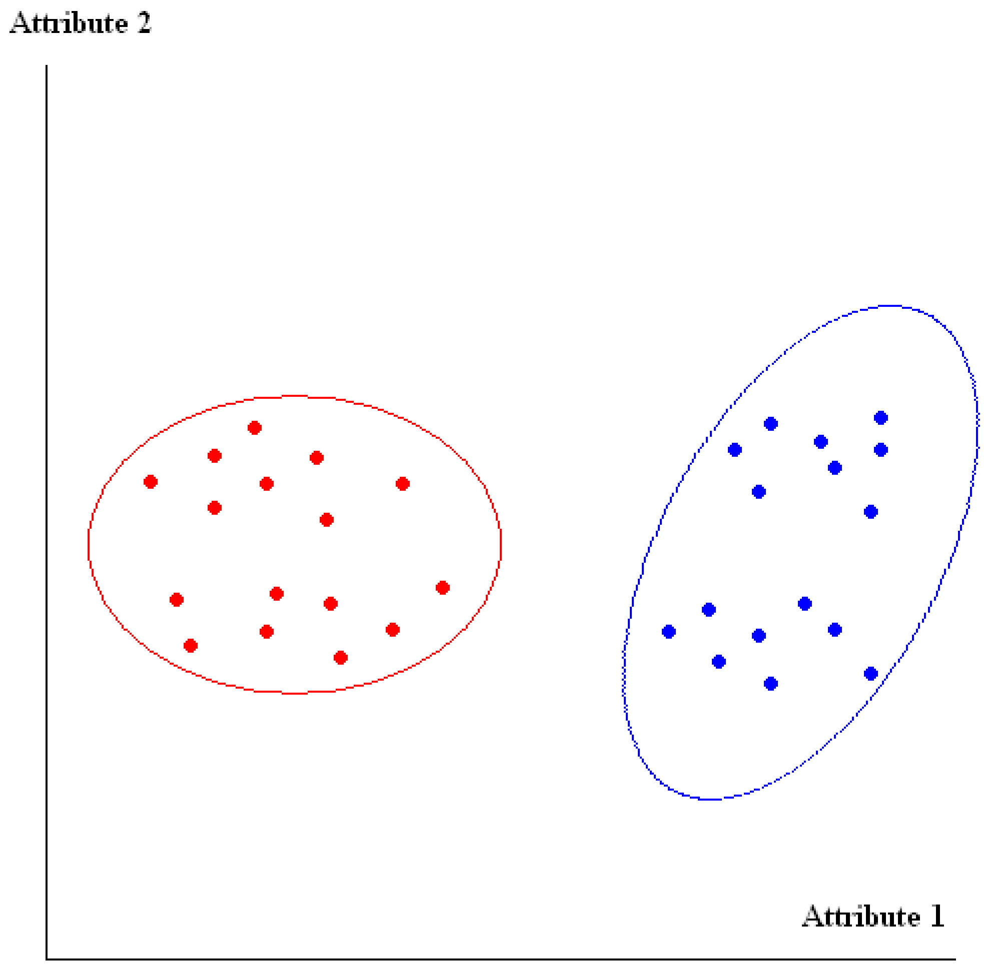
Sample Data Set



Transformed Space



Metric Learning + Clustering



Application: Clustering of Face Poses

Looking to the left

Looking upwards

Picture Courtesy: Clustering with Constraints, S. Basu & I. Davidson

Extensions & pointers

- DistBoost to find a ‘strong’ distance function from a set of ‘weak’ distance functions
  - Weak learner: Fit a mixture of Gaussians under equivalence constraints.
  - Final distance function obtained as a weighted combination of these weak learners.
- Generating constraints
  - Active feedback from user – querying only the most informative instances.
  - Spatial and temporal constraints from video sequences.
  - For content-based image retrieval (CBIR), derived from annotations provided by users.

Curse of Dimensionality

- In many applications, we simply vectorize an image or image patch by a raster-scan.
- 256 x 256 image converts to a 65,536-dimensional vector.
- Images, therefore, are typically very high-dimensional data
- Volume, and hence the number of points required to uniformly sample a space increases exponentially with dimension.
- Affects the convergence of any learning algorithm.
- In some applications, we know that there are only a few variables, for e.g., face pose and illumination.
- Data lie on some low-dimensional subspace/manifold in the high-dimensional space.

Manifold Methods for Vision

- Manifold is a topological space where the local geometry is Euclidean.
- Exist as a part of a higher-dimensional space.
- Some examples:
  - 1-D : line (linear), circle (non-linear)
  - 2-D : 2-D plane (linear), surface of 3-D sphere (non-linear)
- The curse of dimensionality can be mitigated under the manifold assumption.
- Linear dimensionality reduction techniques like PCA have been widely used in the vision community.
- Recent trend is towards non-linear techniques that recover the intrinsic parameterization (pose & illumination).

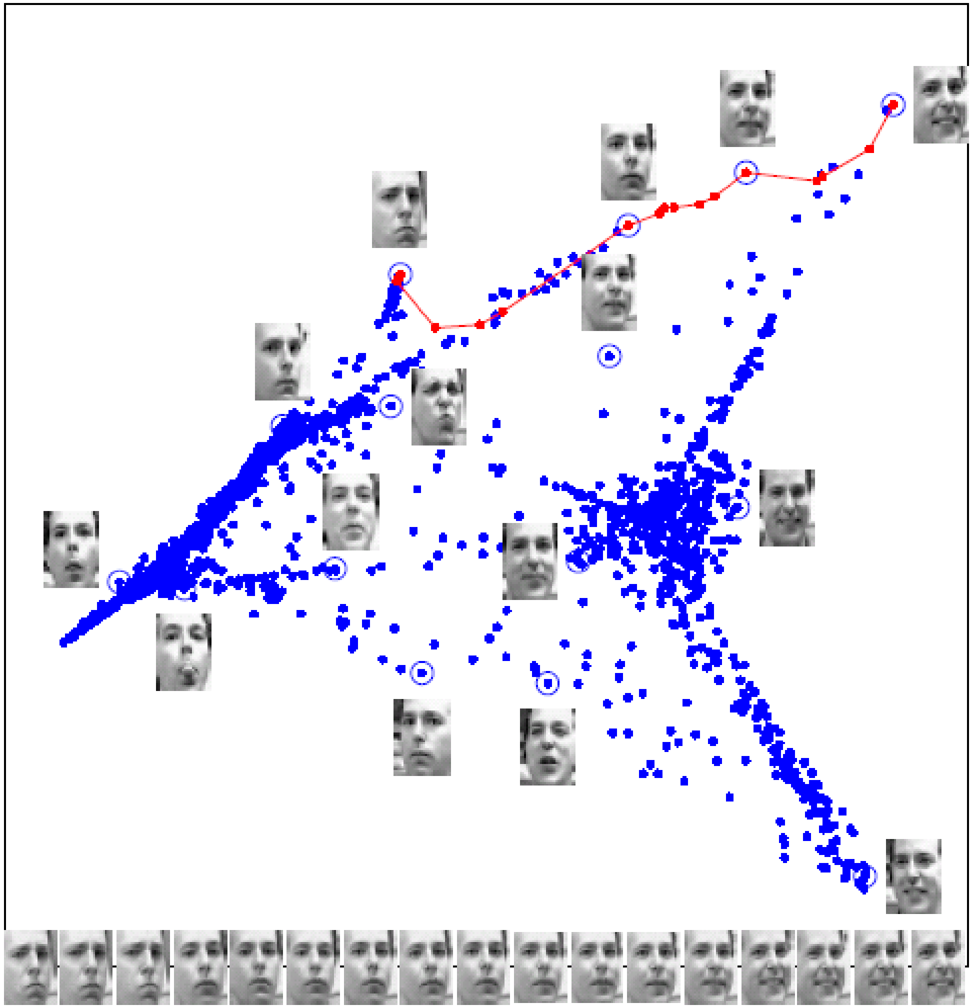
Manifold Embedding Techniques

- Some of the most commonly known manifold embedding techniques:
  - (Kernel) PCA
  - MDS
  - ISOMAP
  - Locally Linear Embedding (LLE)
  - Laplacian Eigenmaps
  - Hessian Eigenmaps
  - Hessian LLE
  - Diffusion Map
  - Local Tangent Space Alignment (LTSA)
- Semi-supervised extensions to many of these algorithms have been proposed.

Manifold Embedding: Basic Idea

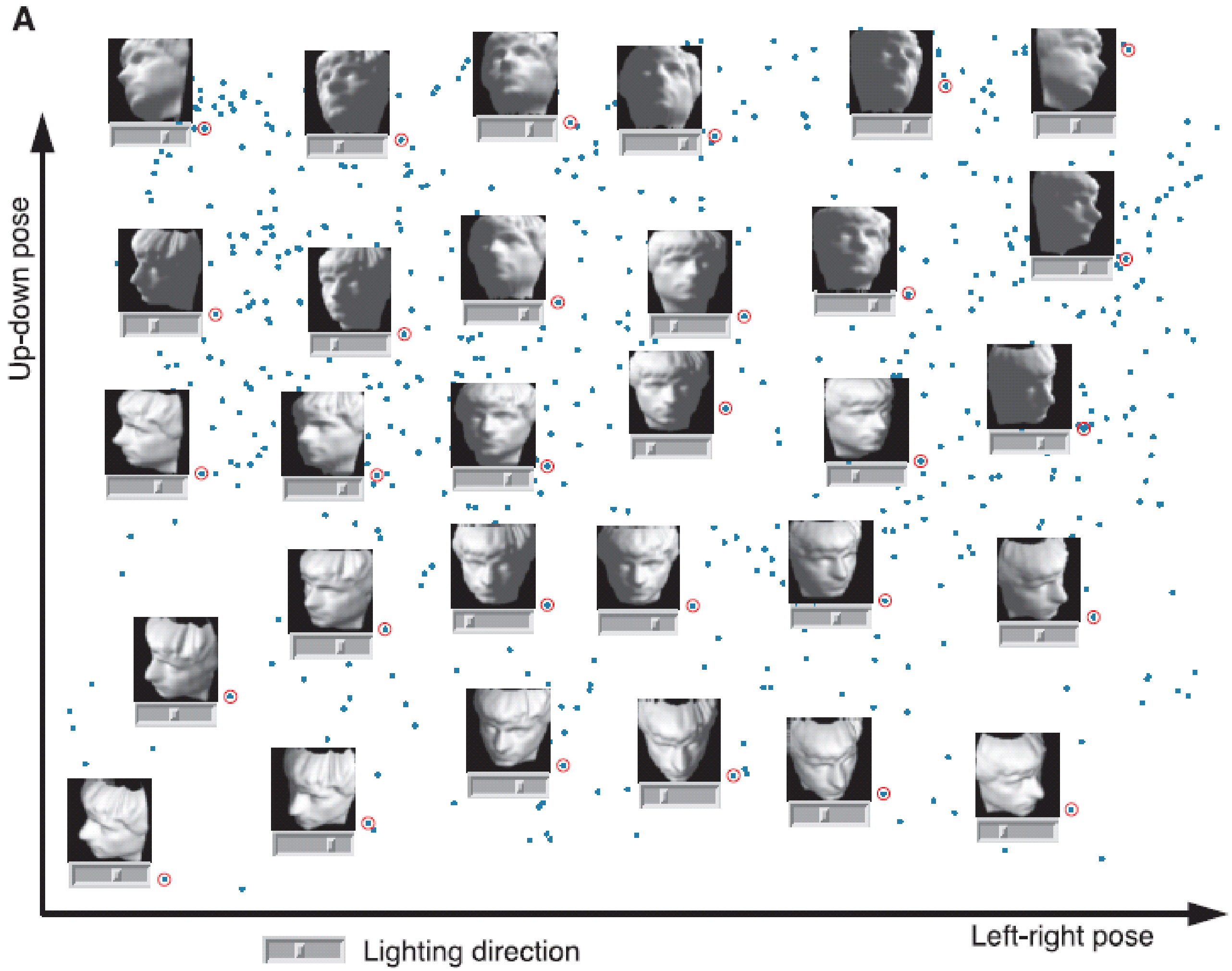
- Most of the manifold methods give a low dimensional embedding, by minimizing a loss function which represents the reconstruction error.
- Almost all of them involve spectral decomposition of a (usually large) matrix.
- Low dimensional embedding obtained represents the intrinsic parameterization recovered from the given data points.
- For e.g., pose, illumination, expression of faces from the CMU PIE Database.
- Other applications include motion segmentation and tracking, shape classification, object recognition.

LLE Embedding



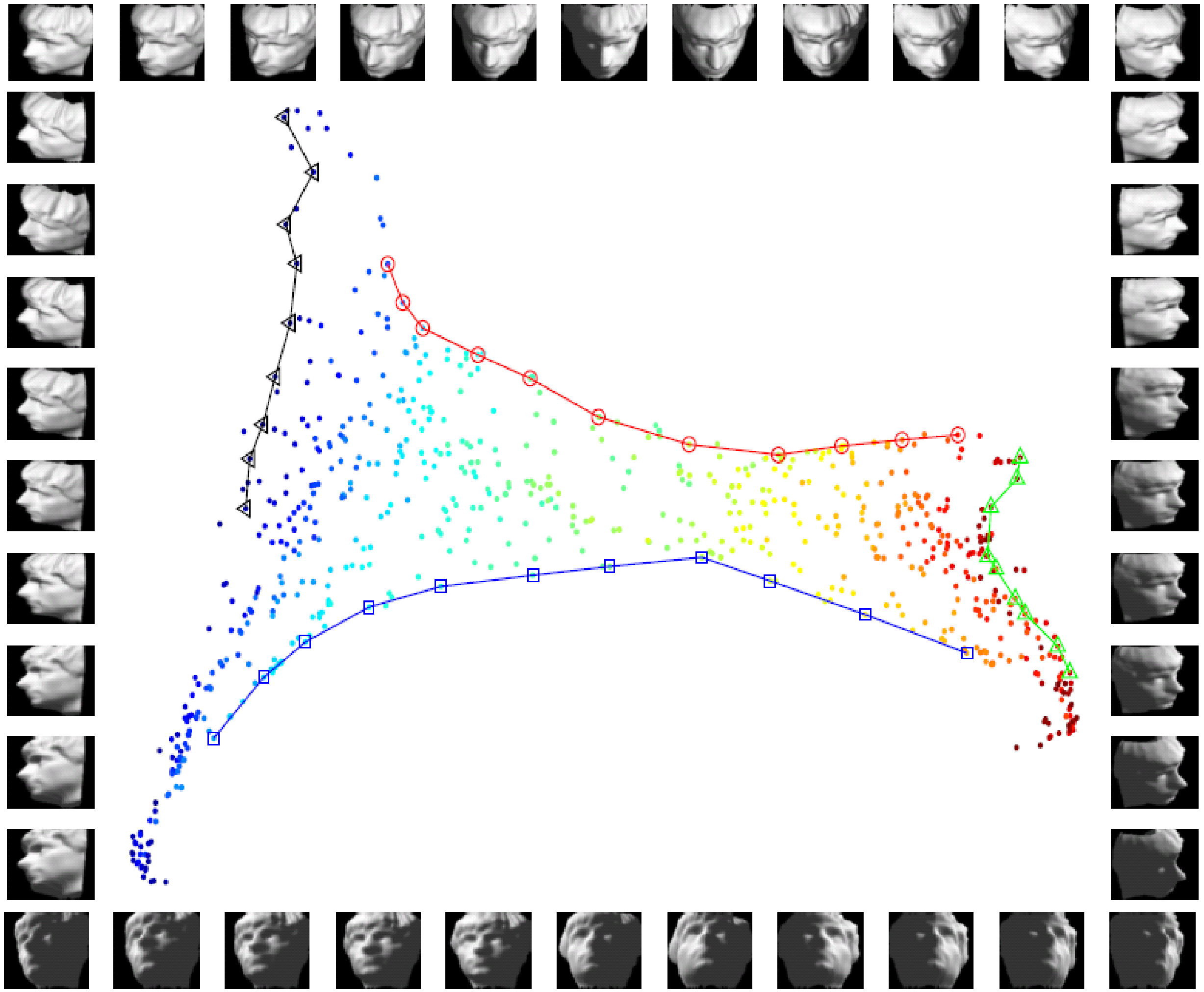
Picture Courtesy: Think Globally, Fit Locally: Unsupervised Learning of Low Dimensional Manifolds (2003) by L.K. Saul & S.T. Roweis

ISOMAP Embedding



Picture Courtesy: A Global Geometric Framework for Nonlinear Dimensionality Reduction by J.B.Tenenbaum, V. de Silva, J. C. Langford in SCIENCE Magazine 2000

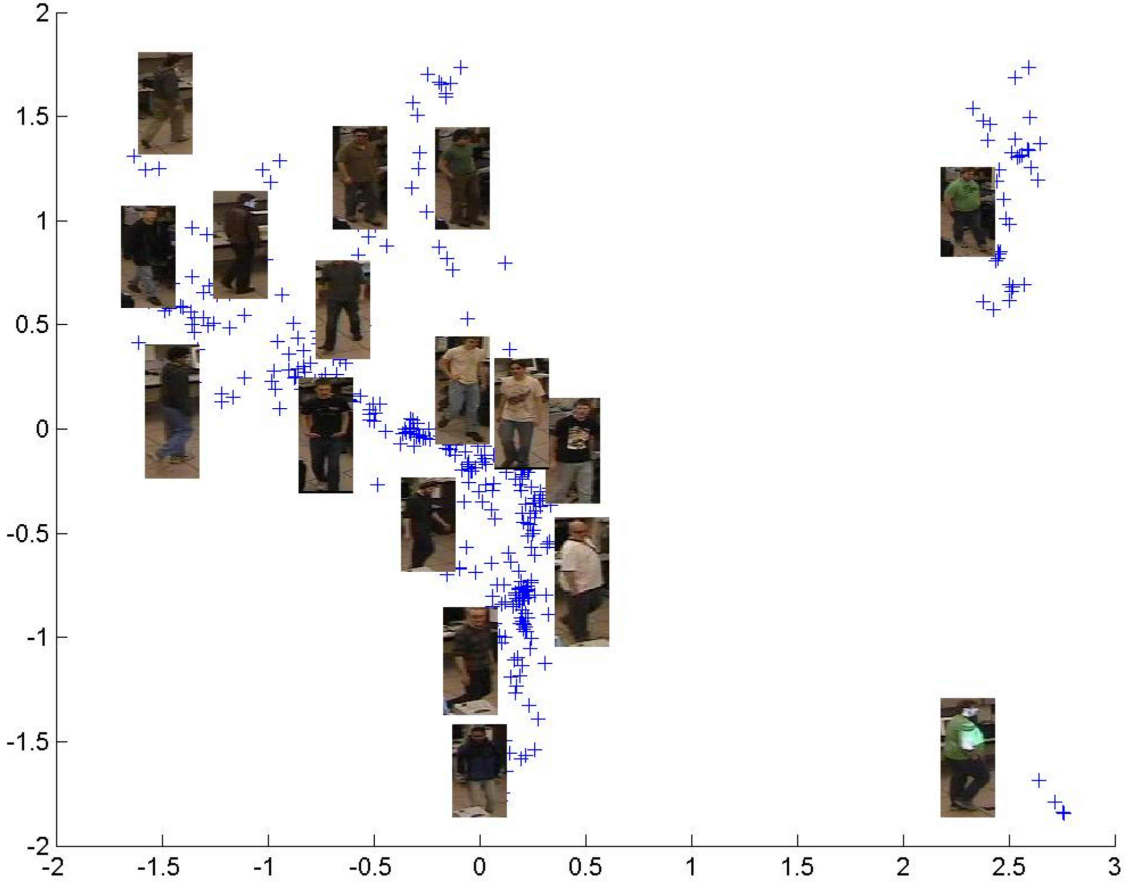
LTSA Embedding



Picture Courtesy: Principal Manifolds and Nonlinear Dimension Reduction via Local Tangent Space Alignment (2002), Z. Zhang & H. Zha

Example: Appearance Clustering

ISOMAP embedding of Region Covariance Descriptors of 17 people.



Sparsity based Learning

- Related to Compressed Sensing.
- Main idea: one can recover certain signals and images from far fewer samples or measurements than traditional methods (Shannon's sampling) use.
- Assumptions:
  - Sparsity: Information rate of a signal is much smaller than suggested by its bandwidth.
  - Incoherence: The original basis in which data exists and the basis in which it is measured are incoherent.

Sparsity based Learning

- Given a large collection of unlabeled images
  - Learn an over complete dictionary from patches of the images using L1 minimization.

$$\min_{b,a} \sum_i \|y^i - \sum_j a^i_j b_j\|_2^2 \; + \; \beta \|a^i\|_1$$

Here vectors y's are vectorized patches of images, b is a matrix constituting the basis vectors of the dictionary and vector a represents the weights of each basis in the dictionary.

- Model the labeled images using this dictionary to obtain sparse weights a.
- Train a classifier/regressor on the a.
- Project the test data onto same dictionary and classification/regression using the learned model.

Example: Altitude estimation of UAV

Given a video of the ground from a down-looking camera on a UAV, can the height of the UAV be estimated?

(0.2m) (0.5m) (0.8m) (1.2m) (1.5m)

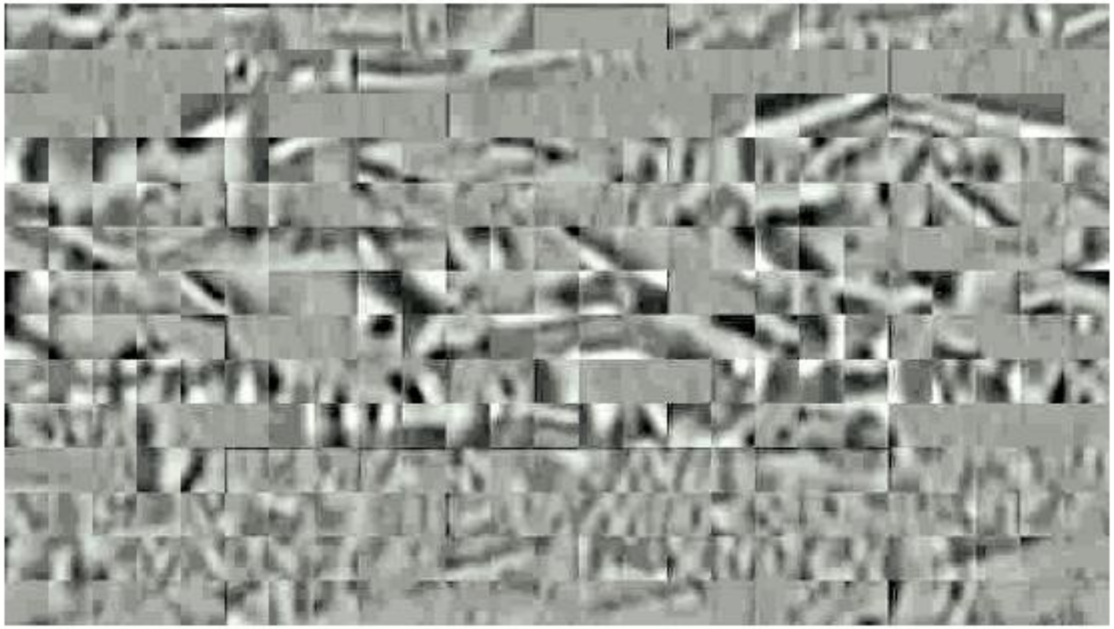
Some sample images of the floor in the lab setting at different heights taken from the base camera of a helicopter.

Altitude estimation continued…

- Arbitrary aerial images from the internet was used to build the dictionary using L1 minimization.

Some sample aerial images used to build the dictionary.

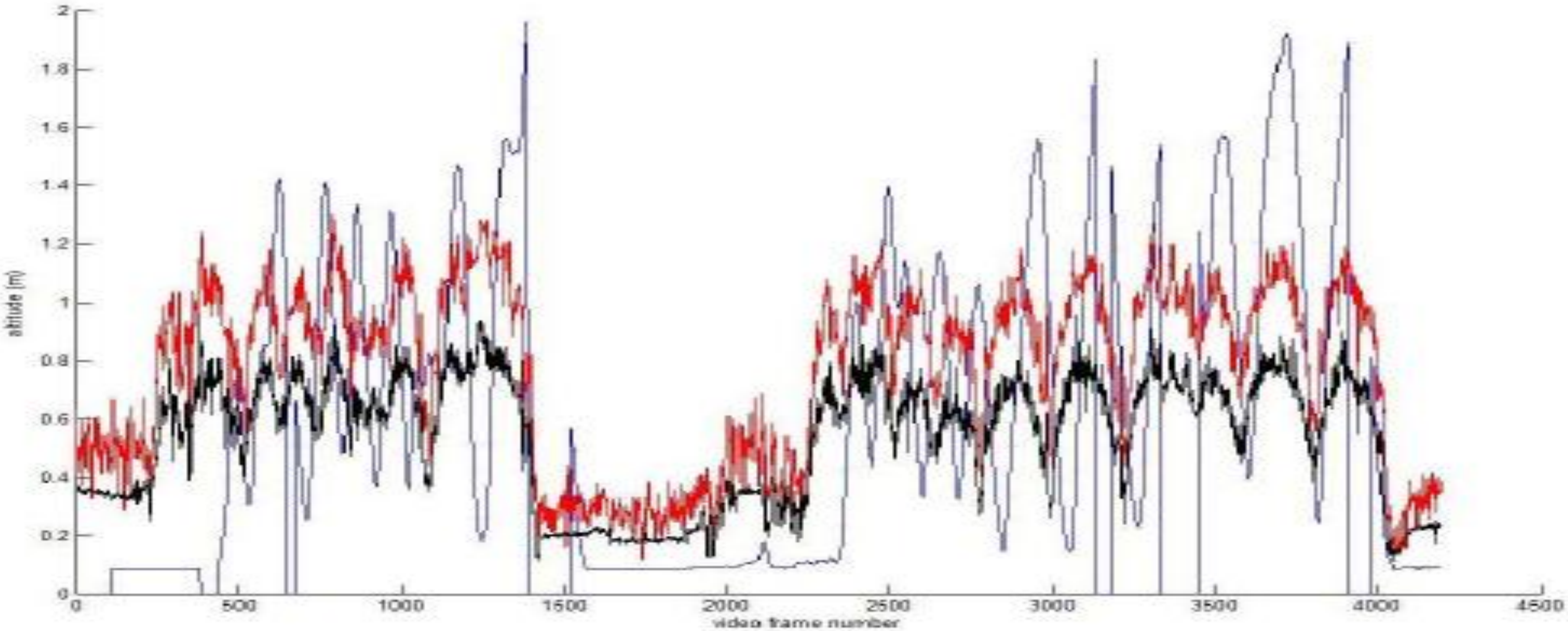
Altitude estimation continued…



350 basis vectors are built using L1 minimization to make the dictionary.

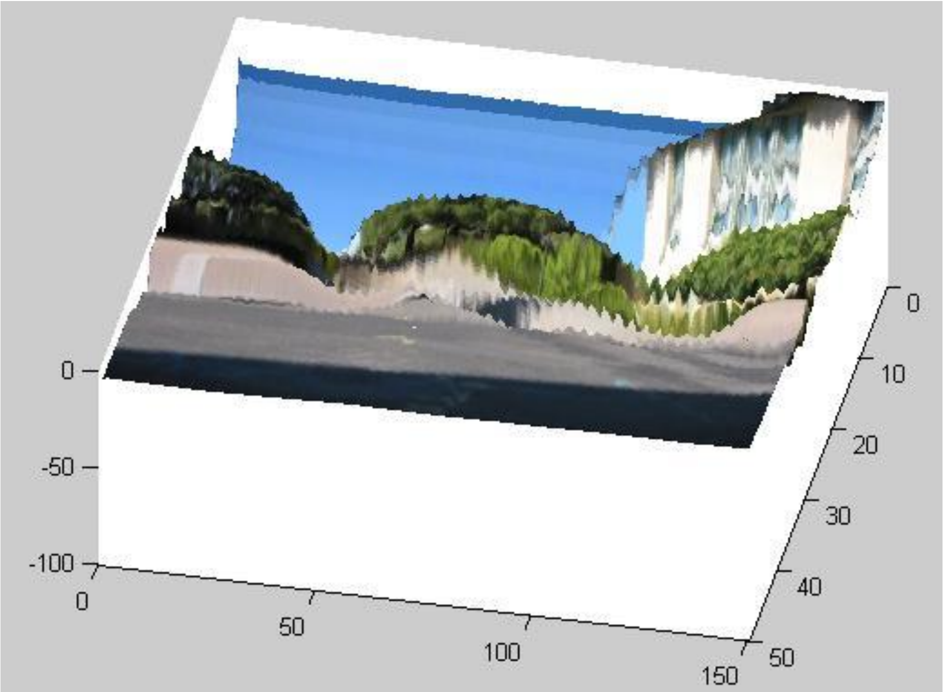
Altitude estimation continued…

- The labeled images shown before are then projected on to this dictionary and an Markov Random Field based regression function is optimized to predict altitudes.
- Some results follow (blue is actual altitude, red is predicted altitude).



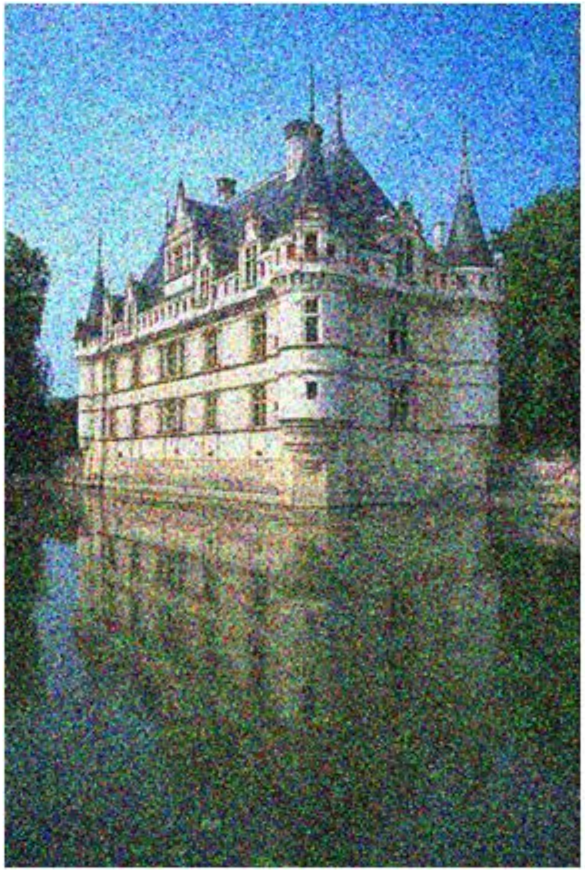
Another Application: 3D reconstruction from a single image

Original image



Reconstructed 3D image

Another Application: Image Denoising



Picture Courtesy: Sparse Representation For Computer Vision and Pattern Recognition (Wright et al, 2009)

Active Learning

- A motivating example: Given an image or a part of it, classify it into a certain category!
- Challenges to be tackled:
  - Large variations in images
  - What is “important” in a given image?
  - Humans are often the judge: very subjective!
- A lot of training is generally required for accurate classification.
- Varied scene conditions like lighting, weather, etc needs further training.

Active Learning

- Basic idea:
  - Traditional supervised learning algorithms passively accept training data.
  - Instead, query for annotations on informative images from the unlabeled data.
  - Theoretical results show that large reductions in training sizes can be obtained with active learning!

But how to find images that are the most informative ?

Active Learning continued…

- One idea uses uncertainty sampling.
- Images on which you are uncertain about classification might be informative!

- What is the notion of uncertainty?
  - Idea: Train a classifier like SVM on the training set.
  - For each unlabeled image, output probabilities indicating class membership.
  - Estimate probabilities can be used to infer uncertainty.
  - A one-vs-one SVM approach can be used to tackle multiple classes.

Active Learning continued…

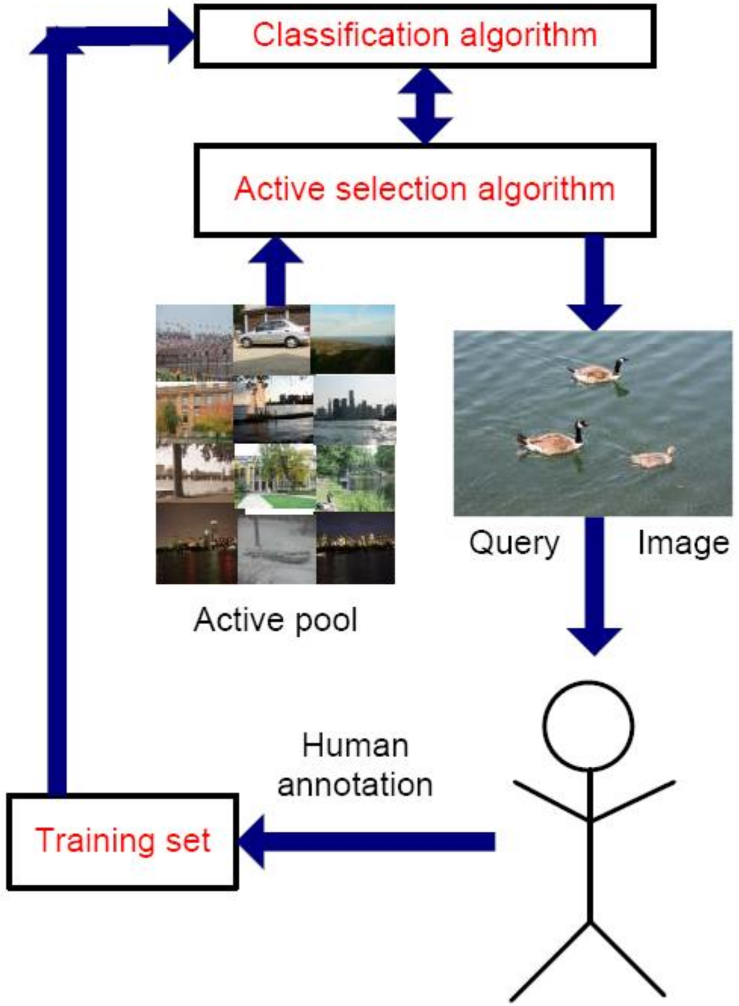
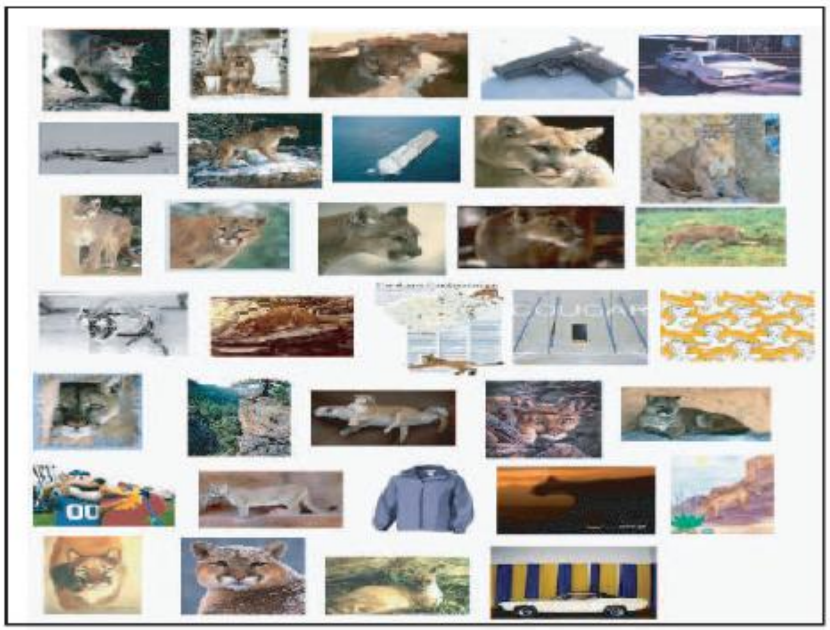


Image Classification using Active Selection

A web search for ‘Cougar’ category



Random selection

Active selection

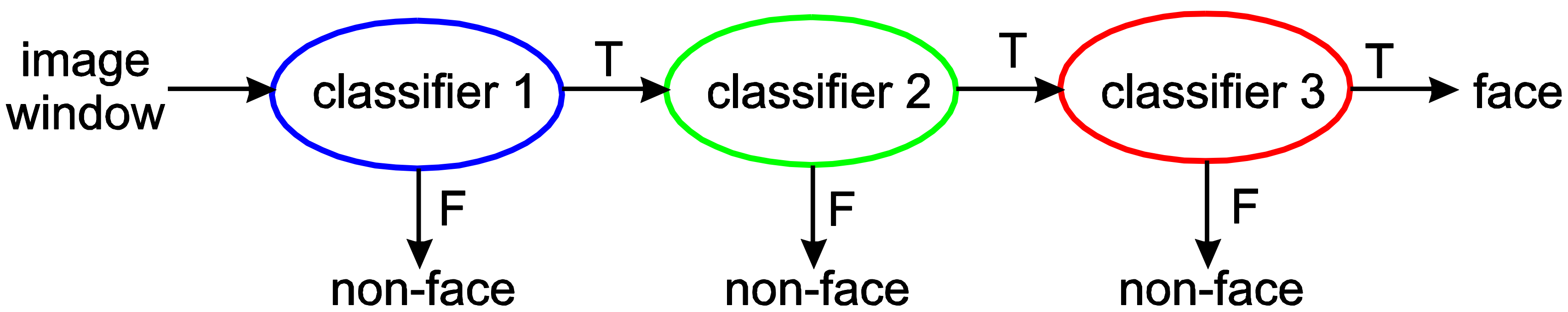
Lesser user input is required in active feedback

Picture courtsey: Entropy based active learning for object categorization, (Holub et al., 2008),

Success stories

Viola-Jones Face Detector (2001)

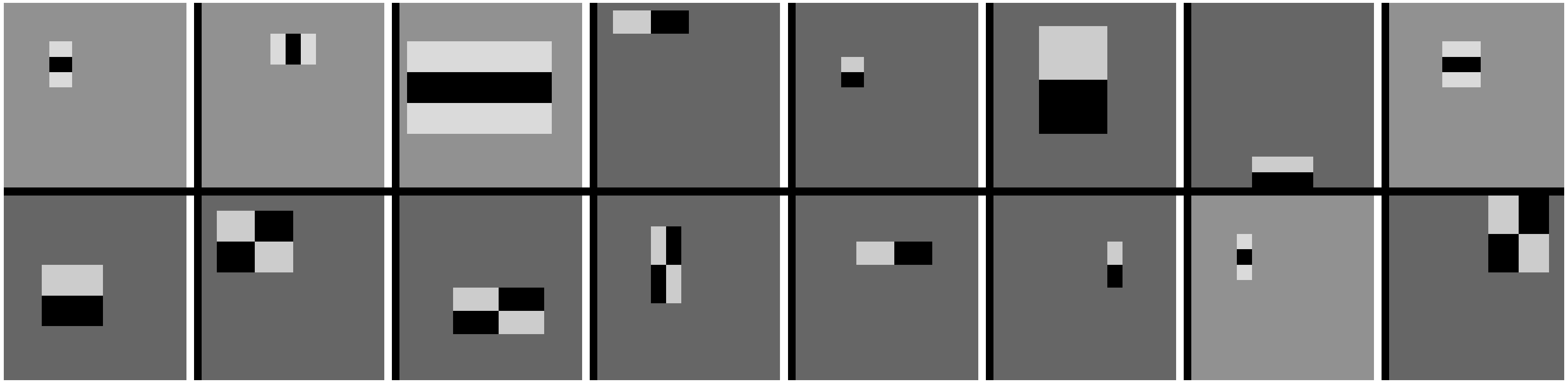
- One of the most notable successes of application of Machine Learning in computer vision.
- World’s first real-time face detection system.
- Available in Intel’s OpenCV library.
- Built as a cascade of boosted classifiers based on the human attentional model.
- Features consist of an over-complete pool of Haar wavelets.



Picture Courtesy: Machine Learning Techniques for Computer Vision (ECCV 2004), C. M. Bishop

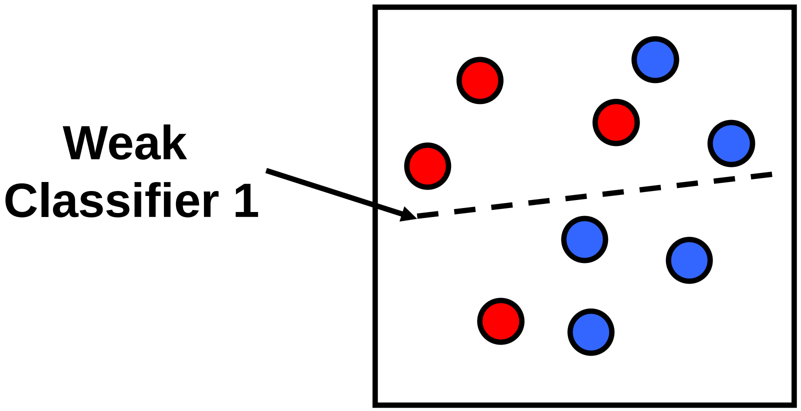
Face Detection

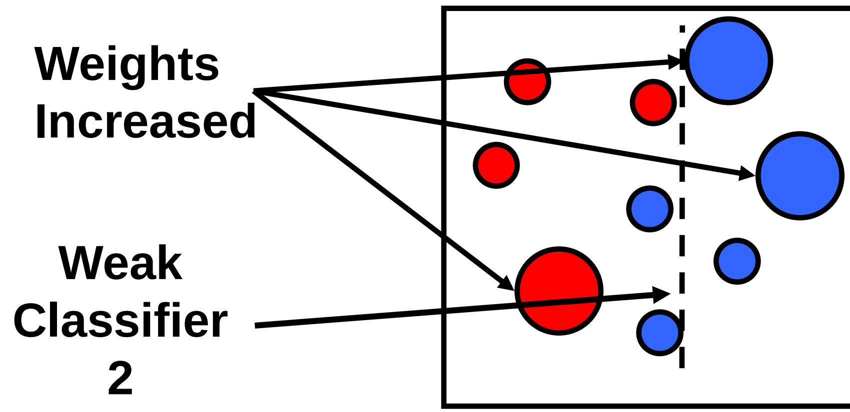
Viola and Jones (2001)

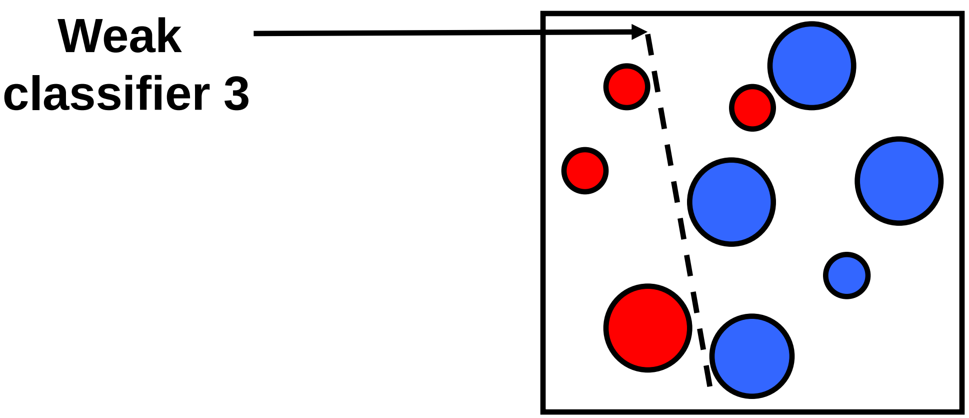


Picture Courtesy: Machine Learning Techniques for Computer Vision (ECCV 2004), C. M. Bishop

Face Detection







Final classifier is linear combination of weak classifiers

Picture Courtesy: Machine Learning Techniques for Computer Vision (ECCV 2004), C. M. Bishop

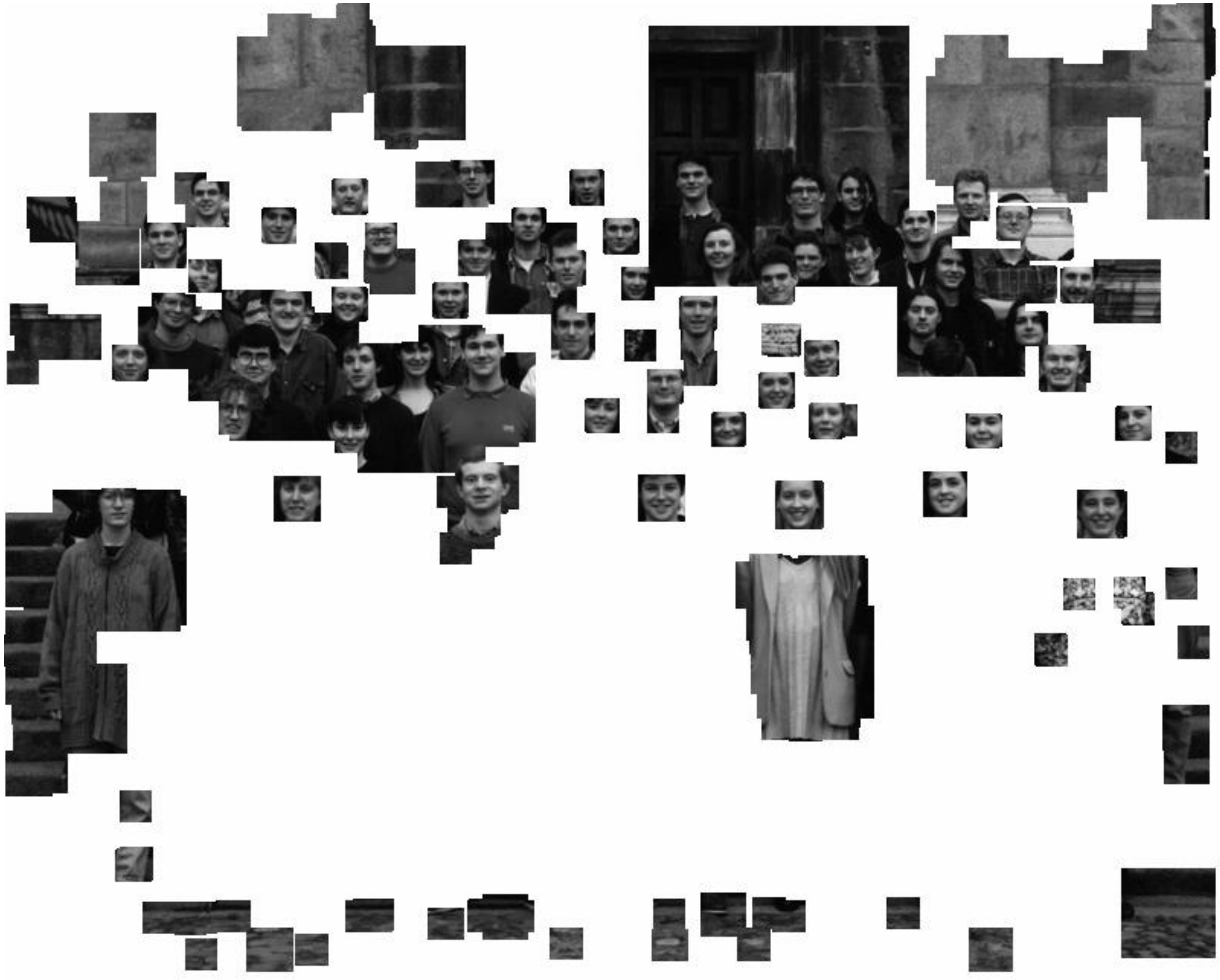
Face Detection

Picture Courtesy: Machine Learning Techniques for Computer Vision (ECCV 2004), C. M. Bishop

Face Detection

Picture Courtesy: Machine Learning Techniques for Computer Vision (ECCV 2004), C. M. Bishop

Face Detection

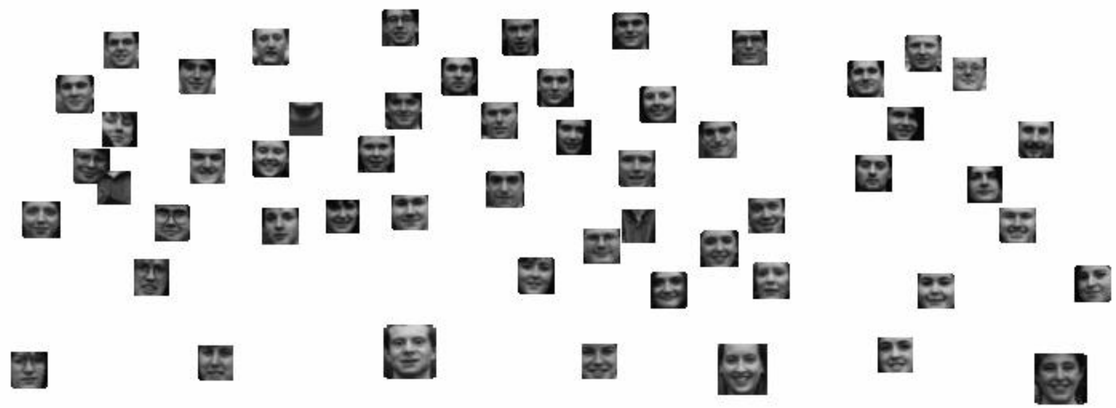


Picture Courtesy: Machine Learning Techniques for Computer Vision (ECCV 2004), C. M. Bishop

Face Detection

Picture Courtesy: Machine Learning Techniques for Computer Vision (ECCV 2004), C. M. Bishop

Face Detection



Picture Courtesy: Machine Learning Techniques for Computer Vision (ECCV 2004), C. M. Bishop

AdaBoost in Vision

Other Uses of AdaBoost

- Human/Pedestrian Detection & Tracking
- Face Expression Recognition
- Iris Recognition
- Action/Gait Recognition
- Vehicle Detection
- License Plate Detection & Recognition
- Traffic Sign Detection & Recognition

Other Features Used in AdaBoost weak classifiers

- Histograms of Oriented Gradients (HOGs)
- Pyramidal HOGs (P-HOGs)
- Shape Context Descriptors
- Region Covariances
- Motion-specific features such as optical flow & other filter outputs

Conclusion: Strengths of ML in Vision

- Solving vision problems through statistical inference
- Intelligence from the crowd/common sense AI (probably)
- Complete autonomy of the computer might not be easily achievable and thus semi-supervised learning might be the right way to go…
- Reducing the constraints over time achieving complete autonomy.

Conclusion: Weakness of ML in Vision

- Application specific algorithms.
- Mathematical intractability of the algorithms leading to approximate solutions.
- Might not work in unforeseen situations.
- Real world problems have too many variables and sensors might be too noisy.
- Computational complexity still the biggest bottleneck for real time applications.

References

[1] A. Singh, R. Nowak, and X. Zhu. **Unlabeled data: Now it helps, now it doesn't**. In Advances in Neural Information Processing Systems (NIPS) 22, 2008.

[2] X. Zhu. **Semi-supervised learning literature survey**. Technical Report 1530, Department of Computer Sciences, University of Wisconsin, Madison, 2005.

[3] Z. Ghahramani, **Unsupervised Learning**, Advanced Lectures on Machine Learning LNAI 3176, Springer-Verlag.

[4] S. Kotsiantis, **Supervised Machine Learning: A Review of Classification Techniques**, Informatica Journal 31 (2007) 249-268

[5] R. Raina, A. Battle, H. Lee, B. Packer, A. Ng, **Self Taught Learning: Transfer learning from unlabeled data**, ICML, 2007.

[6] A. Goldberg, Xi. Zhu, A. Singh, Z. Xu, and R. Nowak. **Multi-manifold semi-supervised learning.** In *Twelfth International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2009.

[7] S. Basu, I. Davidson and K. Wagstaff, ***Constrained Clustering: Advances in Algorithms, Theory, and Applications***, CRC Press, (2008).

[8] B. Settles, **Active Learning Literature Survey**, Computer Sciences Technical report 1648, University of Wisconsin-Madison, 2009.

Thank you!

Slides also available online at:

http://www-users.cs.umn.edu/~cherian/ppt/MachineLearningTut.pdf