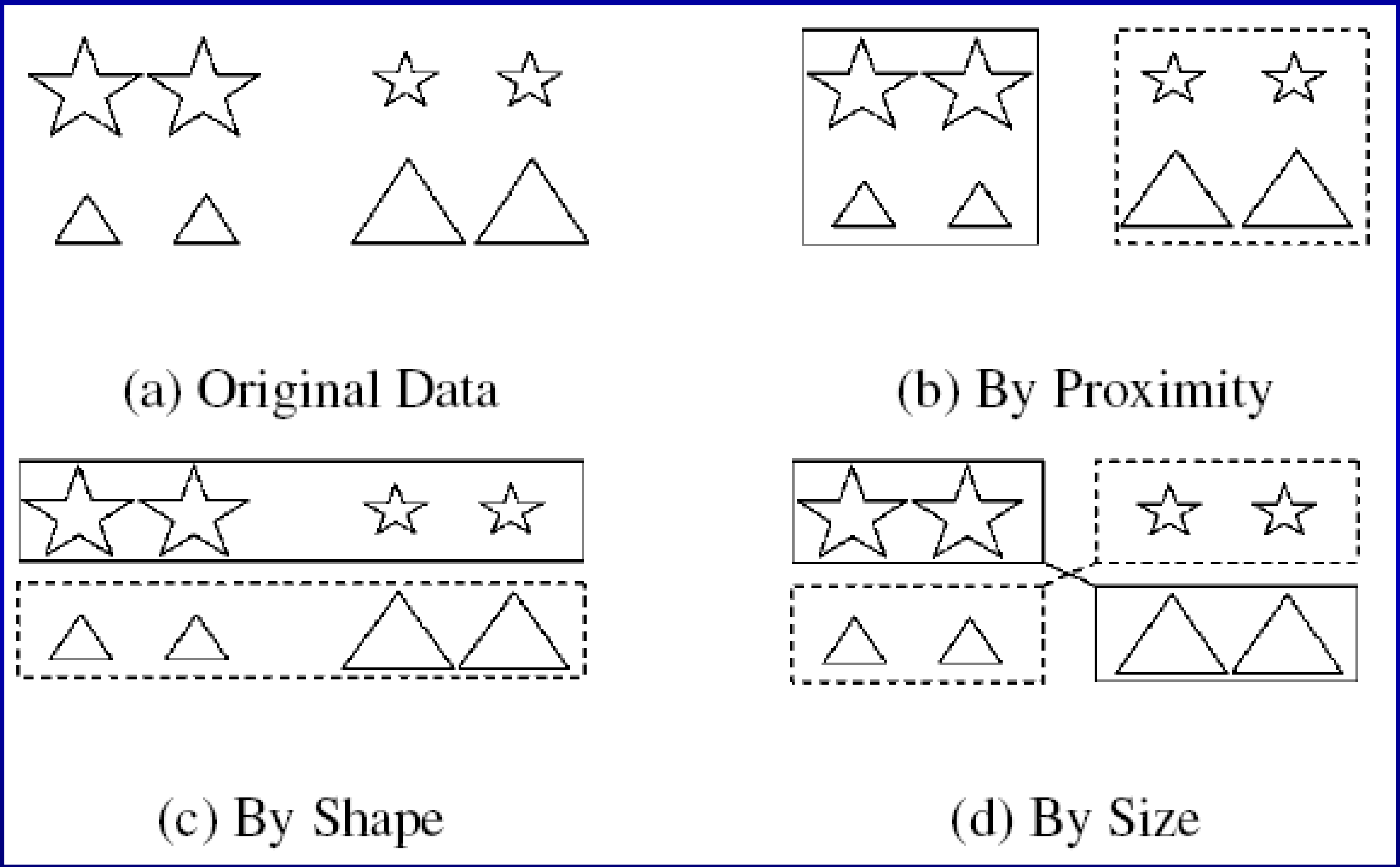
Learning distance functions (demo)

CS 395T: Visual Recognition and Search
April 4, 2008
David Chen

Supervised distance learning

- Learning distance metric from side information
  - Class labels
  - Pairwise constraints
- Keep objects in equivalence constraints close and objects in inequivalence constraints well separated
- Different metrics required for different contexts

Supervised distance learning

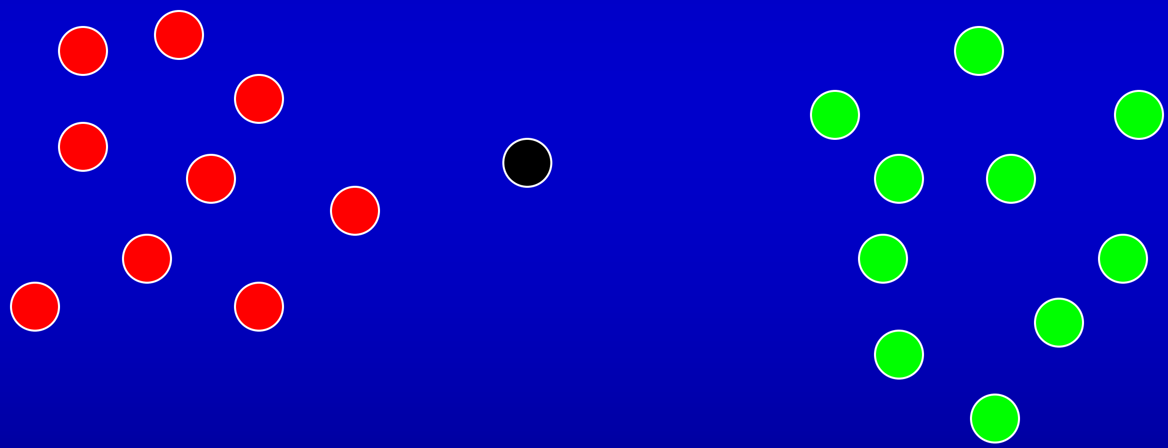


Mahalanobis distance

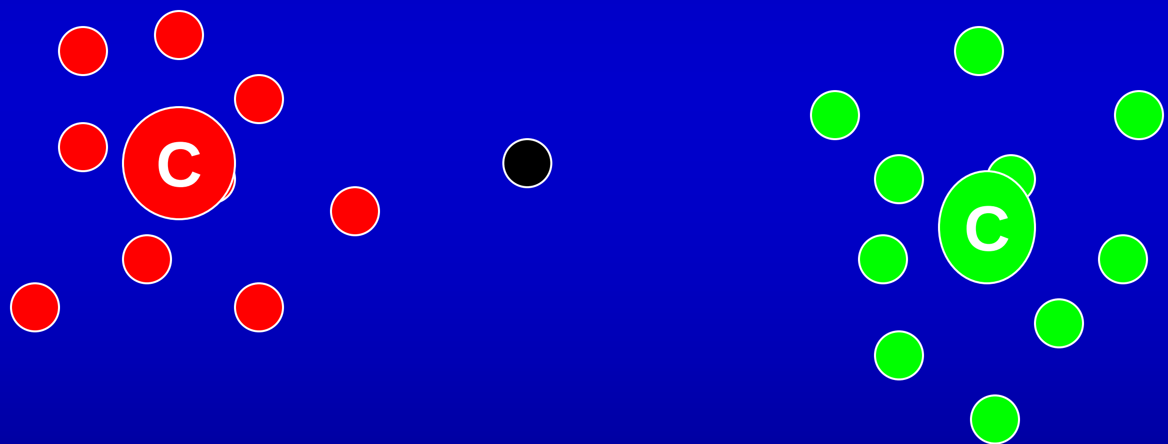
$$d_M(\mathbf{x}_i, \mathbf{x}_j) = \sqrt{(\mathbf{x}_i - \mathbf{x}_j)^\top M (\mathbf{x}_i - \mathbf{x}_j)}$$

- M must be positive semi-definite
- M can be decomposed as $M = A^T A$, where A is a transformation matrix.
- Takes into account the correlations of the data set and is scale-invariant

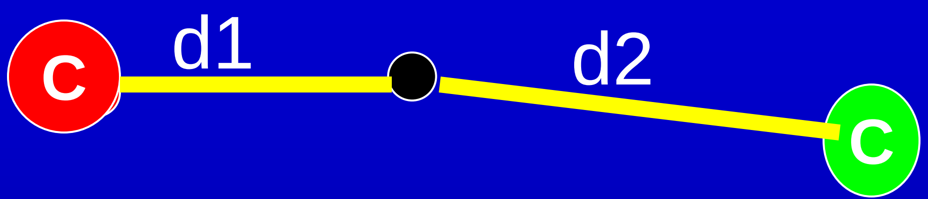
Mahalanobis distance - Intuition



Mahalanobis distance - Intuition



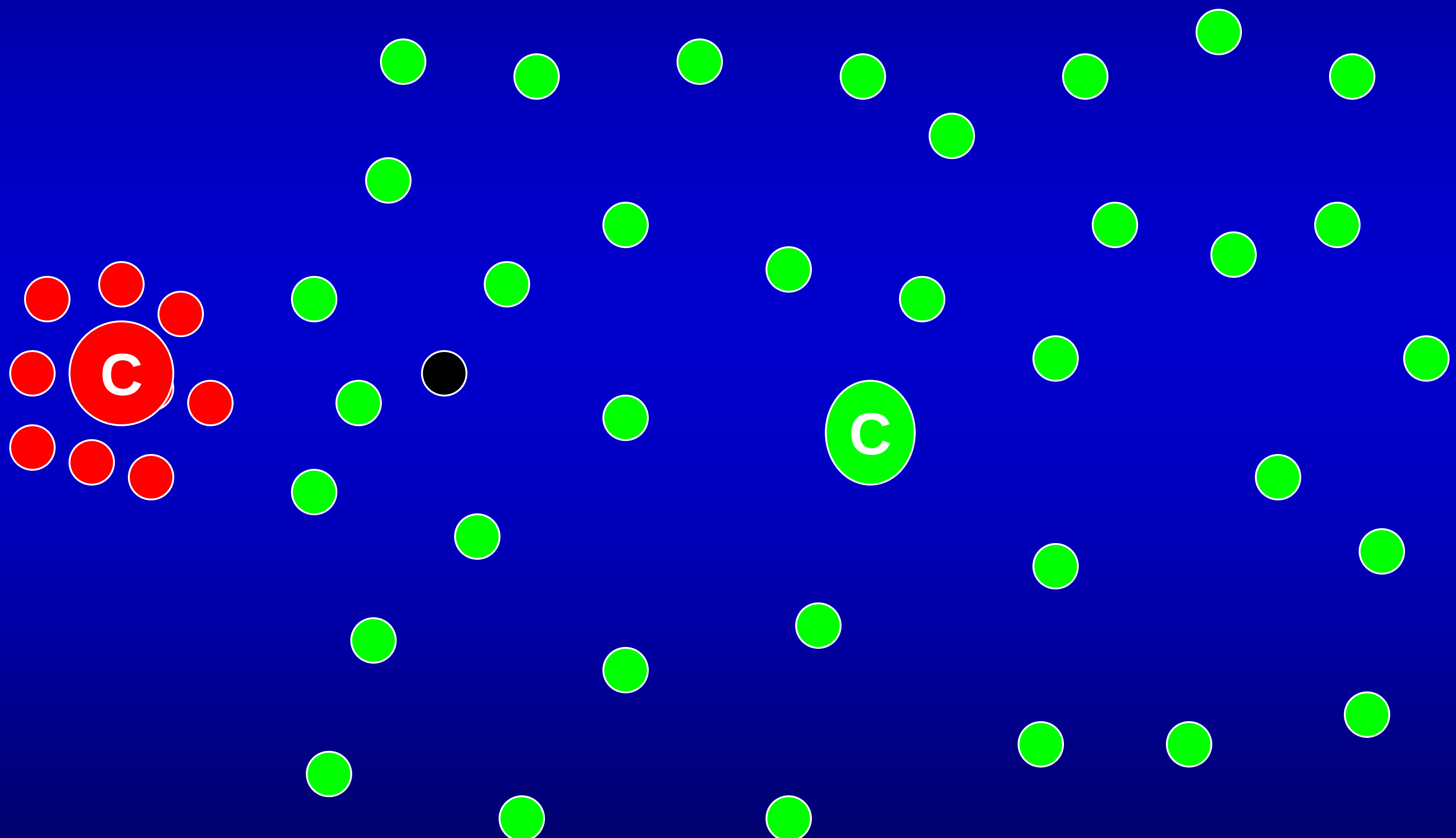
Mahalanobis distance - Intuition



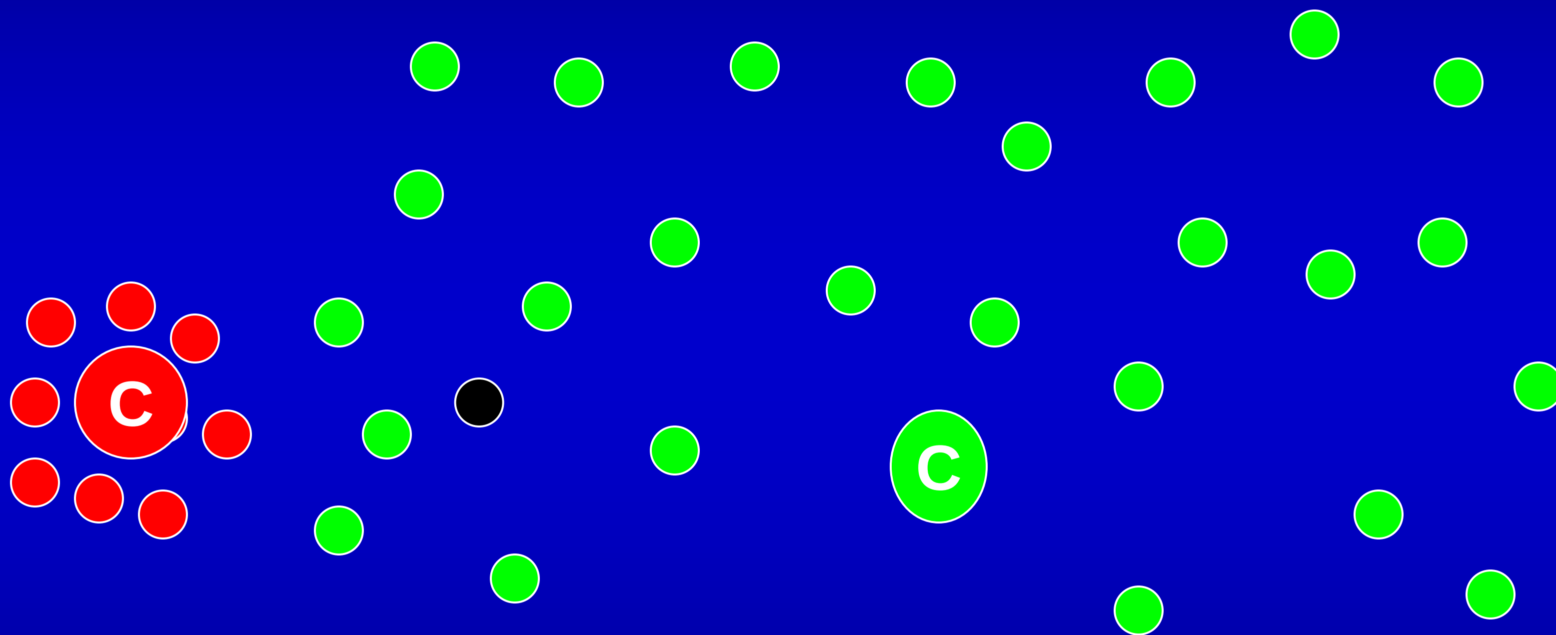
d = |X – C|

d1 < d2 so we classify the point as being red

Mahalanobis distance - Intuition



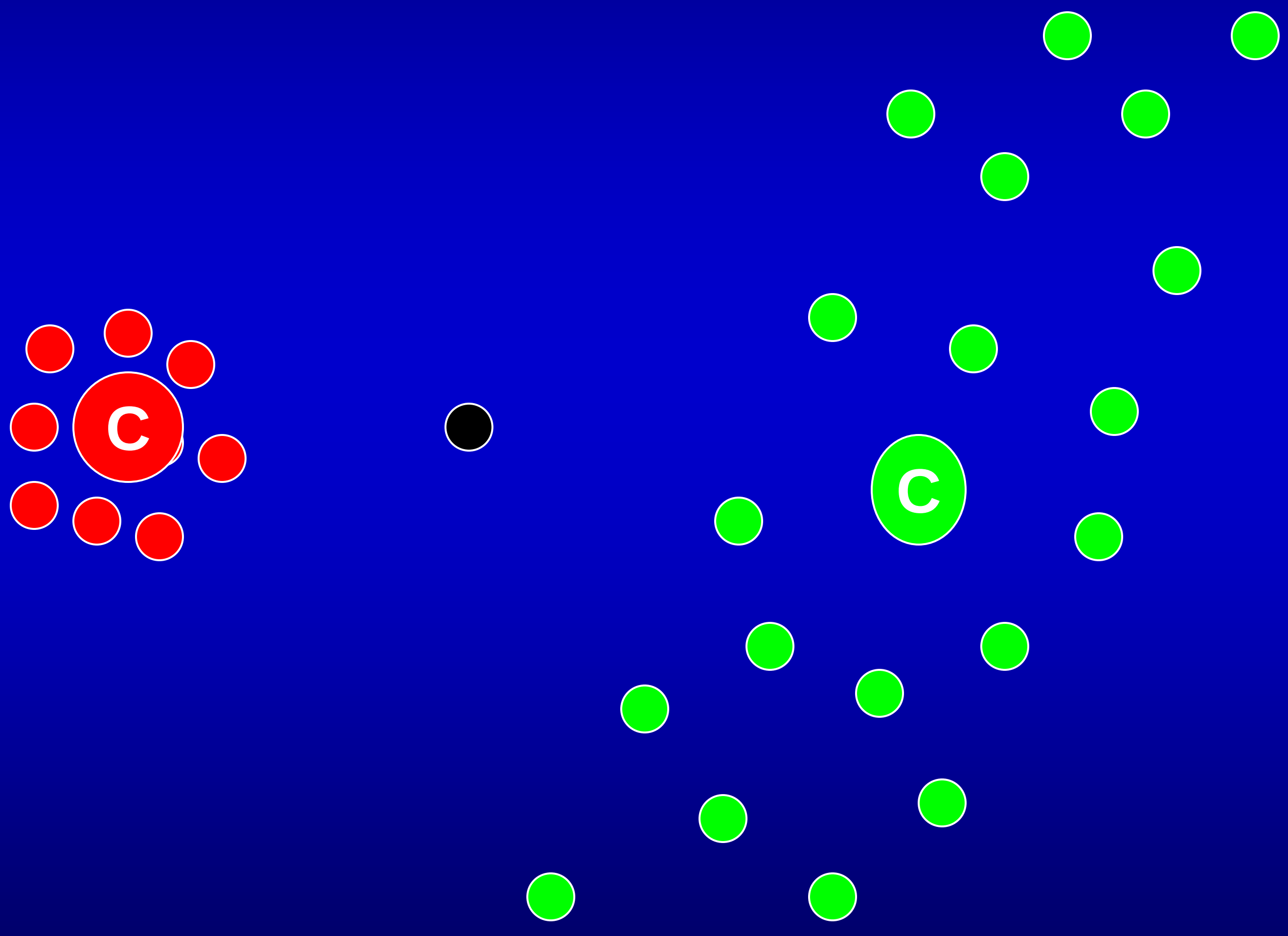
Mahalanobis distance - Intuition



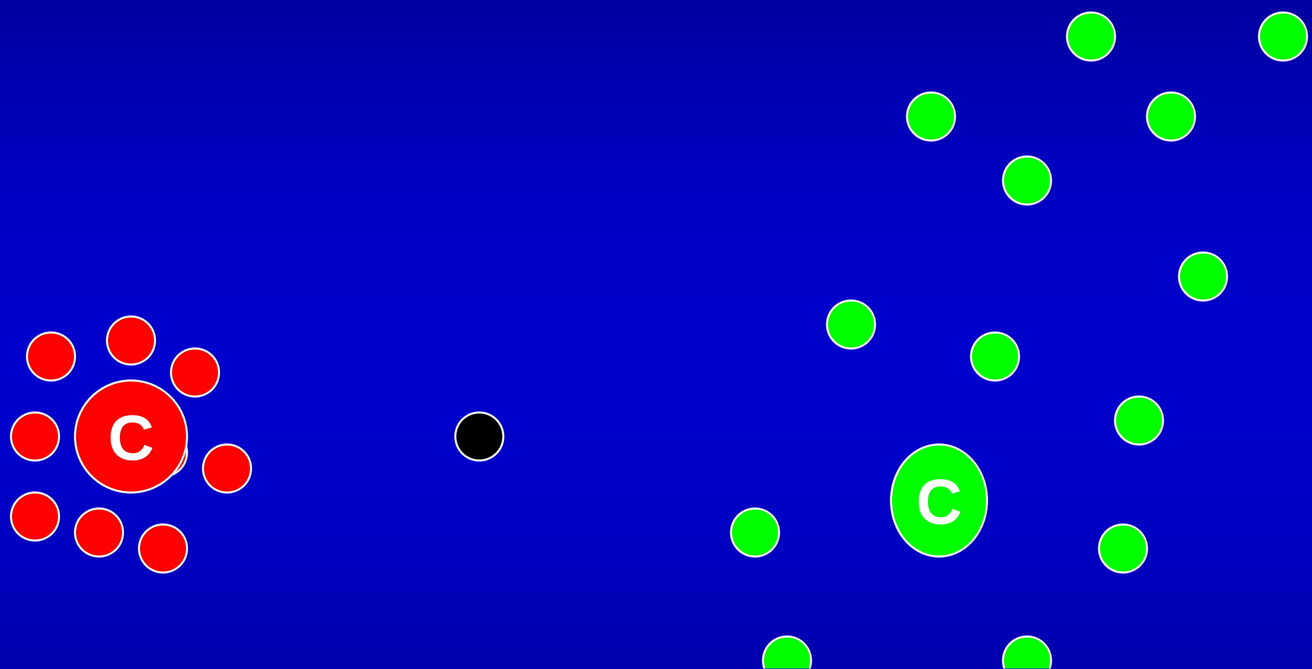
d = |X – C| / std. dev.

So we classify the point as green

Mahalanobis distance - Intuition



Mahalanobis distance - Intuition



Mahalanobis distance is simply |X – C| divided by the width of the ellipsoid in the direction of the test point.

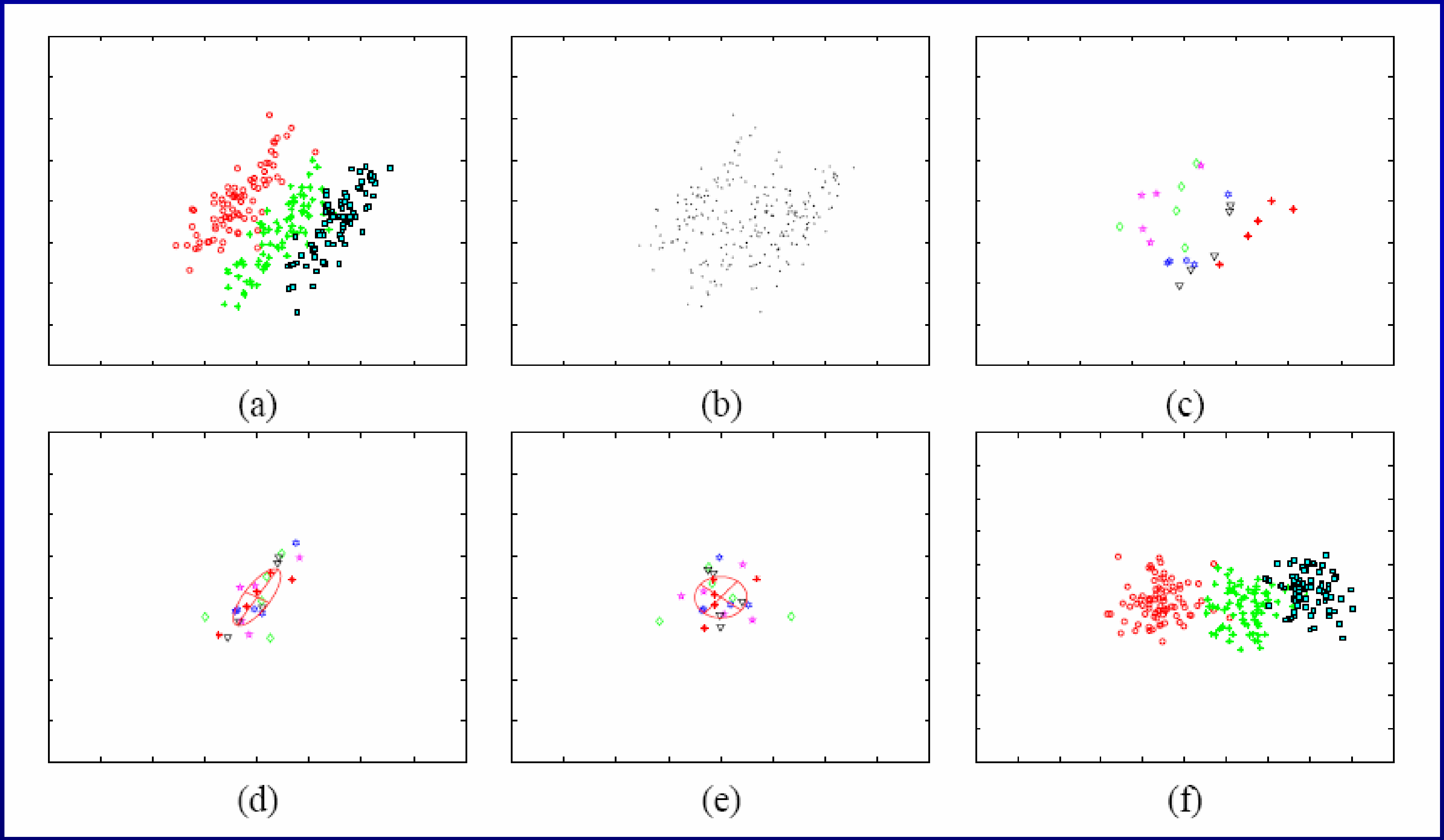
Algorithms

- Relevant Components Analysis (RCA)
- Discriminative Component Analysis (DCA)
- Maximum-Margin Nearest Neighbor (LMNN)
- Information Theoretic Metric Learning (ITML)

Relevant Components Analysis (RCA)

- *Learning a Mahalanobis Metric from Equivalence Constraints* (Bar-Hillel, Hertz, Shental, Weinshall. JMLR 2005)
- Down-scale global unwanted variability within the data
- Uses only positive constraints, or *chunklets*

Relevant Components Analysis (RCA)



Relevant Components Analysis (RCA)

- Given data set $X = \{x_i\}$ for $i = 1{:}N$ and n chunklets $C_j = \{x_{ji}\}$ for $i = 1{:}n_j$
- Compute the within chunklet covariance matrix

$$\hat{C} = \frac{1}{N}\sum_{j=1}^{n}\sum_{i=1}^{n_j}(x_{ji} - m_j)(x_{ji} - m_j)^t$$

- Apply the whitening transformation:

$$\hat{C}:\ W = \hat{C}^{-\frac{1}{2}} \qquad X_{new} = WX$$

- Alternatively $d(x_1, x_2) = (x_1 - x_2)^t \hat{C}^{-1} (x_1 - x_2)$

Relevant Components Analysis (RCA)

Assumptions:

1. The classes have multi-variate normal distributions
2. All the classes share the same covariance matrix
3. The points in each chunklet are an i.i.d. sample from the class

Relevant Components Analysis (RCA)

- Pros
  - Simple and fast
  - Only requires equivalence constraints
  - Maximum likelihood estimation under assumptions
- Cons
  - Doesn’t exploit negative constraints
  - Requires large number of constraints
  - Does poorly when assumptions violated

Discriminative Component Analysis (DCA)

- Learning distance metrics with contextual constraints for image retrieval (Hoi, Liu, Lyu, Ma. CVPR 2006)
- Extension of RCA
- Uses both positive and negative constraints
- Maximize variance between discriminative chunklets and minimize variance within chunklets

Discriminative Component Analysis (DCA)

- Calculate variance of data between chunklets and within chunklets

$$\hat{C}_b = \frac{1}{n_b} \sum_{j=1}^{n} \sum_{i \in D_j} (\mathbf{m}_j - \mathbf{m}_i)(\mathbf{m}_j - \mathbf{m}_i)^\top$$

$$\hat{C}_w = \frac{1}{n} \sum_{j=1}^{n} \frac{1}{n_j} \sum_{i=1}^{n_j} (\mathbf{x}_{ji} - \mathbf{m}_j)(\mathbf{x}_{ji} - \mathbf{m}_j)^\top$$

- Solve this optimization problem

$$J(A) = \arg\max_{A} \frac{|A^\top \hat{C}_b A|}{|A^\top \hat{C}_w A|}$$

Discriminative Component Analysis (DCA)

- Similar to RCA but uses negative constraints
- Slight improvement but faces many of the same issues

Large Margin Nearest Neighbor (LMNN)

- Distance metric learning for large margin nearest neighbor classification (Weinberger, Sha, Zhu, Saul. NIPS 2006)
- K-nearest neighbors should belong to the same class and different classes are separated by a large margin
- Semidefinite programming

Large Margin Nearest Neighbor (LMNN)

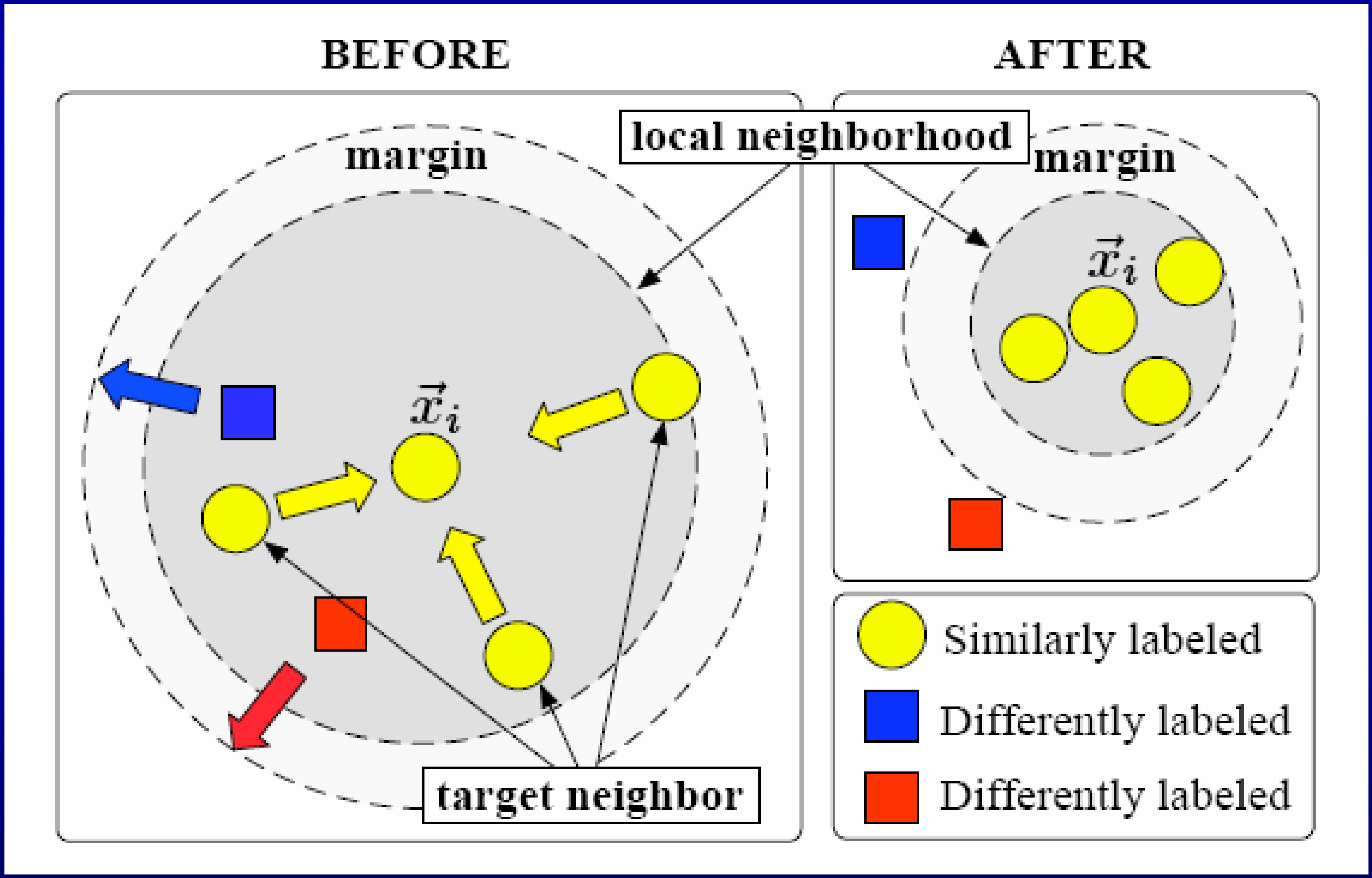
Cost function:

$$\varepsilon(\mathbf{L}) = \sum_{ij} \eta_{ij} \|\mathbf{L}(\vec{x}_i - \vec{x}_j)\|^2 + c \sum_{ijl} \eta_{ij}(1 - y_{il}) \left[1 + \|\mathbf{L}(\vec{x}_i - \vec{x}_j)\|^2 - \|\mathbf{L}(\vec{x}_i - \vec{x}_l)\|^2\right]_+$$

Penalizes large distances between input and its target neighbors

Penalizes small distances between each input and all other inputs that do not share the same label

Large Margin Nearest Neighbor (LMNN)



Large Margin Nearest Neighbor (LMNN)

SDP Formulation:

Minimize $\sum_{ij} \eta_{ij}(\vec{x}_i - \vec{x}_j)^\top \mathbf{M}(\vec{x}_i - \vec{x}_j) + c\sum_{ij} \eta_{ij}(1 - y_{il})\xi_{ijl}$ **subject to:**

(1) $(\vec{x}_i - \vec{x}_l)^\top \mathbf{M}(\vec{x}_i - \vec{x}_l) - (\vec{x}_i - \vec{x}_j)^\top \mathbf{M}(\vec{x}_i - \vec{x}_j) \ \geq\ 1 - \xi_{ijl}$

(2) $\xi_{ijl} \geq 0$

(3) $\mathbf{M} \succeq 0$**.**

Large Margin Nearest Neighbor (LMNN)

- Pros
  - Does not try to keep all similarly labeled examples together
  - Exploits power of kNN classification
  - SDPs: Global optimum can be computed efficiently
- Cons
  - Requires class labels

Extension to LMNN

- An Invariant Large Margin Nearest Neighbor Classifier (Kumar, Torr, Zisserman. ICCV 2007)
- Incorporates invariances
- Adds regularizers

Information Theoretic Metric Learning (ITML)

- Information-theoretic Metric Learning (Davis, Kulis, Jain, Sra, Dhillon. ICML 2007)
- Can incorporate a wide range of constraints
- Regularizes the Mahalanobis matrix A to be close to to a given A_0

Information Theoretic Metric Learning (ITML)

- Cost function:

$$\mathrm{KL}(p(\boldsymbol{x}; A_o) \| p(\boldsymbol{x}; A)) = \int p(\boldsymbol{x}; A_0) \log \frac{p(\boldsymbol{x}; A_0)}{p(\boldsymbol{x}; A)} \, d\boldsymbol{x}$$

- A Mahalanobis distance parameterized by A has a corresponding multivariate Guassian:

$$P(x; A) = 1/Z \exp(-1/2 \, d_A(x, mu))$$

Information Theoretic Metric Learning (ITML)

Optimize cost function given similar and dissimilar constraints

$$\begin{aligned} \min_{A} \quad & \mathrm{KL}(p(\boldsymbol{x}; A_0) \| p(\boldsymbol{x}; A)) \\ \text{subject to} \quad & d_A(\boldsymbol{x}_i, \boldsymbol{x}_j) \leq u \qquad (i,j) \in S, \\ & d_A(\boldsymbol{x}_i, \boldsymbol{x}_j) \geq \ell \qquad (i,j) \in D. \end{aligned}$$

Information Theoretic Metric Learning (ITML)

- Express the problem in terms of the LogDet divergence

$$\begin{aligned}\min_{A\succeq 0,\boldsymbol{\xi}} \quad & D_{\ell\mathrm{d}}(A, A_0) + \gamma \cdot D_{\ell\mathrm{d}}(\mathrm{diag}(\boldsymbol{\xi}), \mathrm{diag}(\boldsymbol{\xi}_0)) \\ \text{s. t.} \quad & \mathrm{tr}(A(\boldsymbol{x}_i - \boldsymbol{x}_j)(\boldsymbol{x}_i - \boldsymbol{x}_j)^T) \le \xi_{c(i,j)} \quad (i,j) \in S, \\ & \mathrm{tr}(A(\boldsymbol{x}_i - \boldsymbol{x}_j)(\boldsymbol{x}_i - \boldsymbol{x}_j)^T) \ge \xi_{c(i,j)} \quad (i,j) \in D,\end{aligned}$$

- Optimized in O(cd^2) time
  - c: number of constraints
  - d: dimension of data
  - Learning Low-rank Kernel Matrices. (Kulis, Sustik, Dhillon. ICML 2006)

Information Theoretic Metric Learning (ITML)

- Flexible constraints
  - Similarity or dissimilarity
  - Relations between pairs of distances
  - Prior information regarding the distance function
- No computation of eigenvalue or semi-definite programming

UCI Dataset

- UCI Machine Learning Repository
- Asuncion, A. & Newman, D.J. (2007). UCI Machine Learning Repository [http://www.ics.uci.edu/~mlearn/MLRepository.html]. Irvine, CA: University of California, School of Information and Computer Science.

UCI Dataset

	# Instances	# Features	# Classes
Iris	150	4	3
Wine	178	13	3
Balance	625	4	3
Segmentation	210	19	7
Pendigits	10992	16	10
Madelon	2600	500	2

Methodology

- 5 runs of 10-fold cross validation for Iris, Wine, Balance, Segmentation
- 2 runs of 3-fold cross validation for Pendigits and Madelon
- Measures accuracy of kNN classifier using the learned metric
  - K = 3
- All possible constraints used except for ITML and Pendigits

UCI Results

	L2	RCA	DCA	LMNN	ITML
Iris	96.00	**96.67**	**96.67**	95.60	96.53
Wine	71.01	**98.88**	**98.88**	97.08	93.71
Balance	79.97	79.62	79.58	82.50	**89.06**
Segmentation	76.29	20.19	20.57	**86.86**	82.48
Pendigits	99.27	**99.37**	**99.37**	99.16	99.26
Madelon	**69.83**	51.21	51.21	63.92	**69.83**

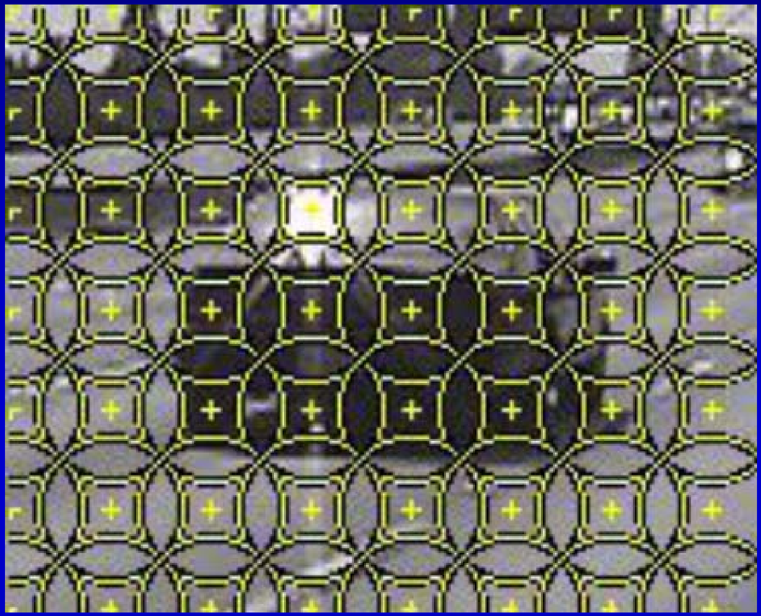
Pascal Dataset

- Pascal VOC 2005

	Motorbikes	Bicycles	People	Cars
Training	214	114	84	272
Test (test 1)	216	114	84	275

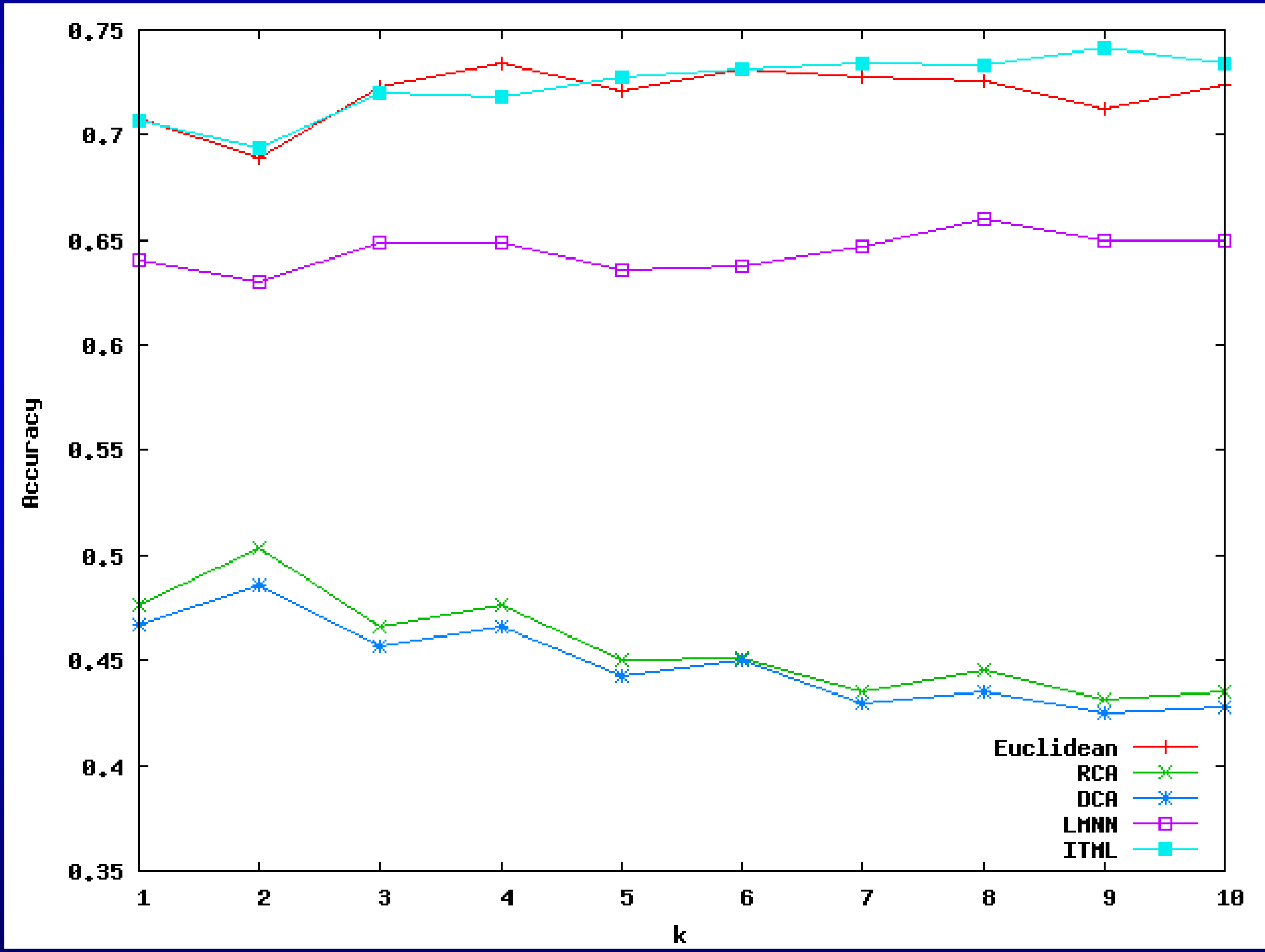
- Using Xin's large overlapping features and visual words (200)
- Each image represented as a histogram of the visual words

Pascal Dataset

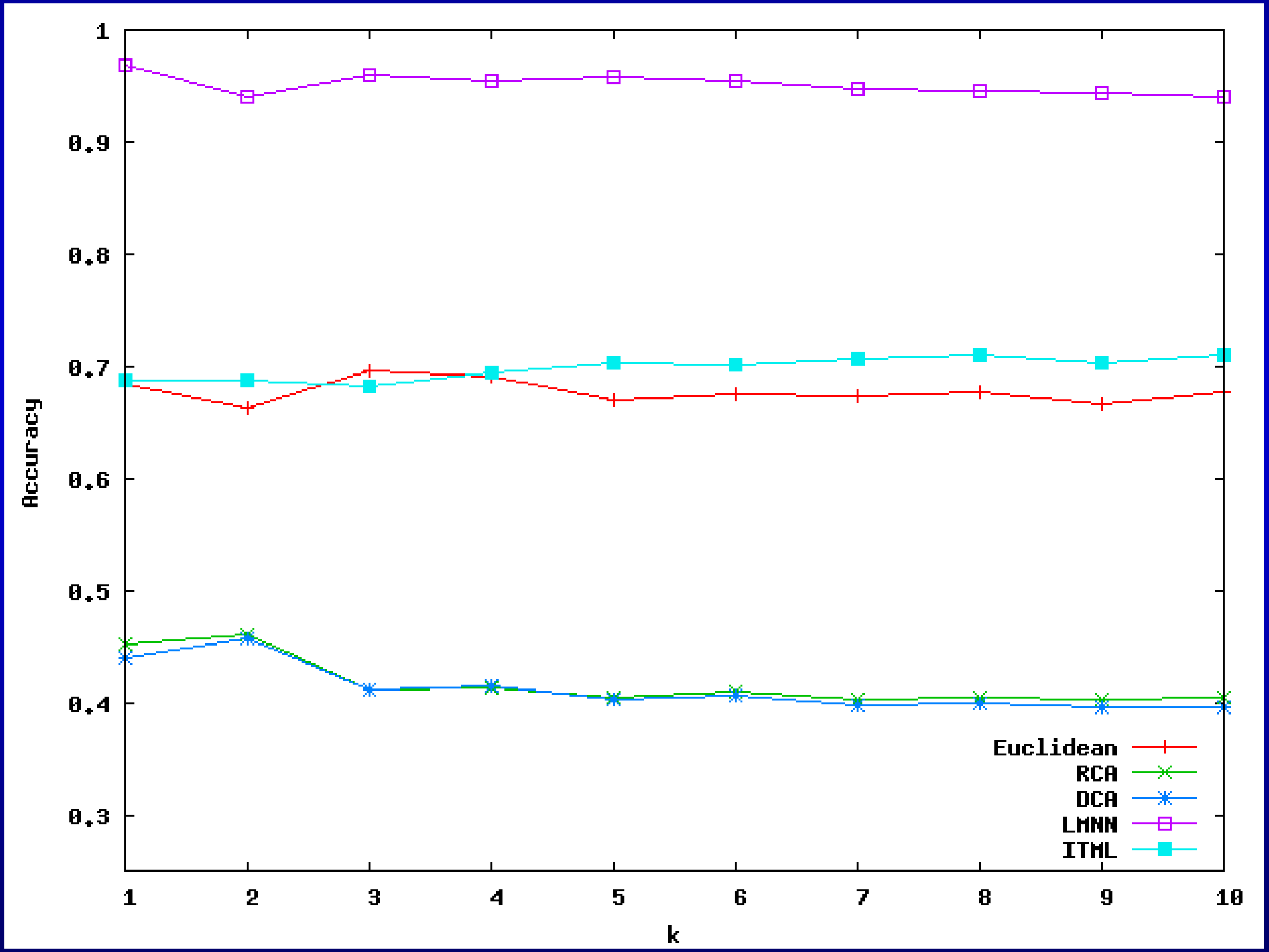


- SIFT descriptors for each patch
- K-means to cluster the descriptors into 200 visual words

Results (test set)



Results (training set)



Results

L2

RCA

DCA

LMNN

ITML

Results

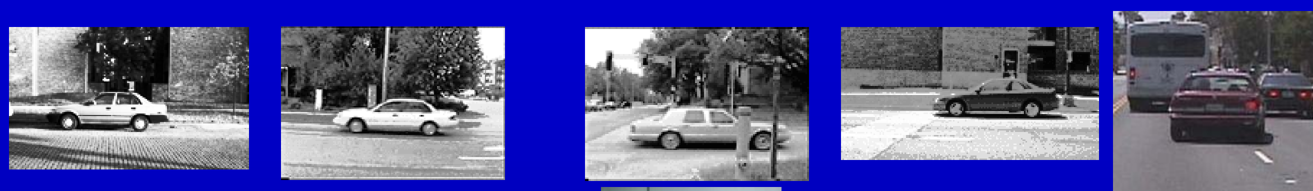
- L2
- RCA
- DCA
- LMNN
- ITML

Results

L2

RCA

DCA



LMNN

ITML

Results

L2

RCA

DCA

LMNN

ITML

Discussion

- Matches a lot of background due to uniform sampling
- Metric learning does not replace good feature construction
- Using PCA to first reduce the dimensionality might help
- Try Kernel versions of the algorithms

Tools used

- DistLearnKit, Liu Yang, Rong Jin
  - http://www.cse.msu.edu/~yangliu1/distlearn.htm
  - Distance Metric Learning: A Comprehensive Survey, by L. Yang, Michigan State University, 2006
- ITML, Jason V. Davis and Brian Kulis and Prateek Jain and Suvrit Sra and Inderjit S. Dhillon
  - http://www.cs.utexas.edu/users/pjain/itml/
  - Information-theoretic Metric Learning (Davis, Kulis, Jain, Sra, Dhillon. ICML 2007)