Discriminative k-metrics

Arthur Szlam and Guillermo Sapiro

k q flats

The k q-flats algorithm (Kambhatla and Leen, many others) is a generalization of k-means....

- Input: n points X in $\mathbb{R}^d$ ($X \subset \mathbb{R}^{d \times n}$), numbers k, q.
- Output: subsets $P_1, ..., P_k$ and q flats $B_1, ..., B_k$ minimizing

$$E(P_1, ..., P_k) = \sum_{i=1}^{k} \sum_{x \in P_i} ||x - x_{B_i}||^2 \qquad (1)$$

- x_{B_i} is the projection of the point x onto the q dimensional best fit plane B_i passing through P_i.

k q-flats, dictionary design, and sparse matrix factorization

$$\sum_{i=1}^{k} \sum_{x \in P_i} ||x - x_{B_i}||^2$$

with the B_i constrained to pass through the origin is the same as

$$\arg\min_{B \in \mathbb{R}^{d \times kq}, C \in \mathbb{R}^{kq \times n}} ||BC - X||_{\text{Fro}},$$

$$||C_j||_0 \leq q,$$

with C constrained to have a block structure.

i.e. we find a basis B of kq vectors and coefficients C so that $BC \sim X$.

$$B = d\{ \big(\overbrace{B_1}^{q} \quad \overbrace{B_2}^{q} \quad ... \quad \overbrace{B_K}^{q} \big)$$

$$C = \begin{array}{c} \\ q\{ \\ q\{ \\ q\{ \\ q\{ \\ q\{ \end{array} \begin{array}{c} \overbrace{}^{|P_1|} \quad \overbrace{}^{|P_2|} \quad \quad \quad \overbrace{}^{|P_K|} \\ \begin{pmatrix} C_1 & 0 & 0 & \cdots & 0 \\ 0 & C_2 & 0 & \cdots & 0 \\ 0 & 0 & C_3 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & 0 \\ 0 & 0 & 0 & 0 & C_K \end{pmatrix} \end{array}$$

Connection with "manifold models" of data:

If the neighborhood of each data point is well represented by the tangent plane at the data point, we expect there to be a good set of planes.

might as well find the best one with respect to reconstruction error...

k q flats for supervised learning

- choose k and q,
- find best k q-planes approximating the data points of each class.
- For each new data point, find its closest plane in each class. Label by closest plane.

i.e parameterize the "manifold" describing class 1, 2, etc; assign unlabeled points to their closest manifold.

That is: find B_{ij} and P_{ij} minimizing

$$E(\mathcal{P}, \mathcal{B}) = -\sum_{i=1}^{m} \sum_{j=1}^{k} \sum_{x \in P_{ij}} ||B_{ij}x||^2$$

- $P_{i,.}$ is a partition of class i
- $\mathcal{P} = \{P_{ij}\}$ is the collection of all these sets,
- B_{ij} is the set of basis vectors associated to P_{ij},
- $\mathcal{B} = \{B_{ij}\}$ is the collection of all these vectors.

Problem: sparse representation strategies as presented not ideal for discriminative learning problems. Only learn representations, not differences!

Solution: change the energy (and generalize planes).

Discriminative k metrics energy

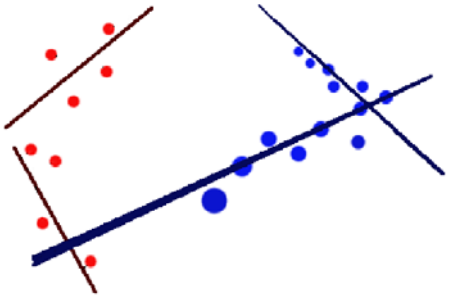
$$E(\mathcal{P},\mathcal{B}) = \sum_{i=1}^{m}\sum_{j=1}^{k}\left[\sum_{x\in P_{ij}} g_1(||B_{ij}x||^2) + \sum_{x\notin C_i} g_2(||B_{ij}x||^2)\right]$$

- $P_{i,.}$ is a partition of class i
- $\mathcal{P} = \{P_{ij}\}$ is the collection of all these sets,
- B_{ij} is the set of basis vectors associated to P_{ij},
- $\mathcal{B} = \{B_{ij}\}$ is the collection of all these vectors. $||B_{ij}x||$ is the norm of x in a **Mahalanobis metric**. Thus we represent our data by sets of metrics, instead of sets of flats.

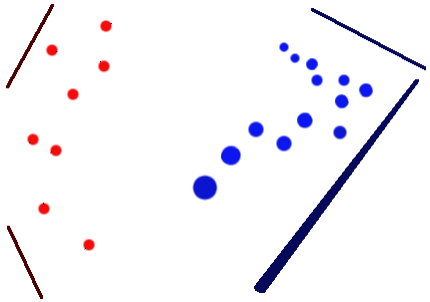
$$E(\mathcal{P}, \mathcal{B}) = -\sum_{i=1}^{m} \sum_{j=1}^{k} \sum_{x \in P_{ij}} ||B_{ij}x||^2$$

vs.

$$\sum_{i=1}^{m} \sum_{j=1}^{k} \left[\sum_{x \in P_{ij}} g_1(||B_{ij}x||^2) + \sum_{x \notin C_i} g_2(||B_{ij}x||^2) \right],$$



Representational



Discriminative

for g_i we take

$$g_1(z) = \alpha_1 \left[(\mu_1 - z)_+\right]^2,$$

and

$$g_2(z) = \alpha_2 \left[(z - \mu_2)_+\right]^2,$$

where

$$a_+ := \begin{cases} a & a > 0 \\ 0 & \text{otherwise.} \end{cases}$$

The μ are margins- parameters we choose by cross validation. α we fix as $\alpha_1 = 2$ and $\alpha_2 = 1$

To minimize the new energy, we use a stochastic gradient descent. The gradient of the energy with respect to a metric F_{ij} is given by:

$$\frac{\partial}{\partial B_{ij}} \left[\sum_{x \in P_{ij}} g_1(||B_{ij}x||^2) + \sum_{x \notin C_i} g_2(||F_{ij}x||^2) \right]$$

$$= \sum_{x \in P_{ij}} 2g_1'(||B_{ij}x||^2) B_{ij} x x^T + \sum_{x \notin C_i} 2g_2'(||B_i x||^2) B_{ij} x x^T.$$

To further improve results, we use the “1/3 trick”, i.e multiple restarts, classify with majority vote.

- MNIST digits: 70000 28×28 images of handwritten digits divided into 60000 training examples and 10000 test examples. Preprocessed by projection onto first 50 principal components.

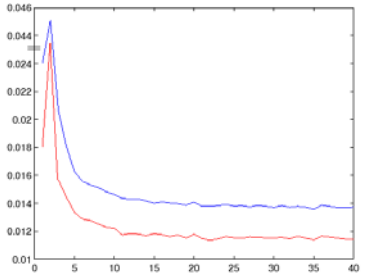
- The 20-newsgroups dataset, consisting of 18477 documents from one of 20 newsgroups represented by its binary term document matrix, with a vocabulary of the 5000 most common words (as in Larochelle and Bengio 2008). The data is divided into the standard 7505/11269 test/train split.

- The ISOLET datset, consiting of 200 speakers saying each letter of the alphabet twice. 617 audio features have been extracted from each sample. The data is divided into a standard training set of the first 150 speakers, and a test set of the last 50.

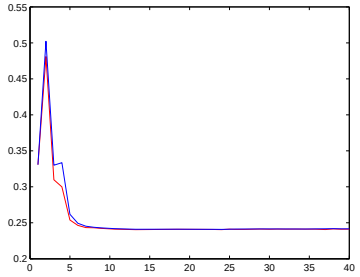
	MNIST	20 newsgroups	Isolet
k q-metrics (m)	1.15	24.0	3.4
k q-metrics	1.36	24.0	3.5
k q-flats	1.6	33.1	4.3
SVM	1.4	30	3.3

Errors in percent

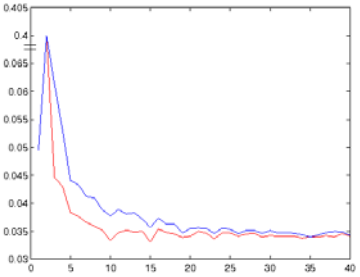
Timings: error rate versus passes through the data



MNIST



20 news.

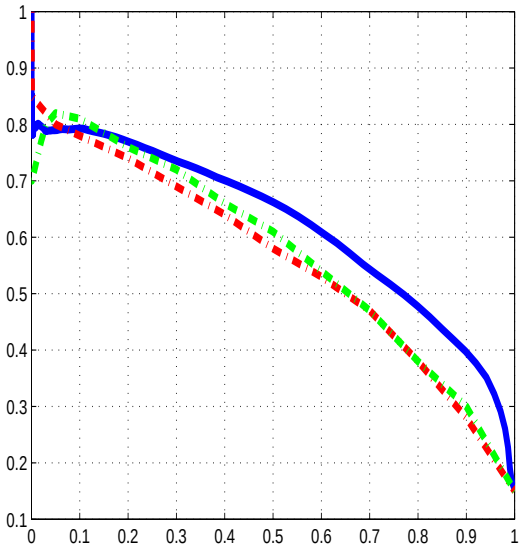


isolet

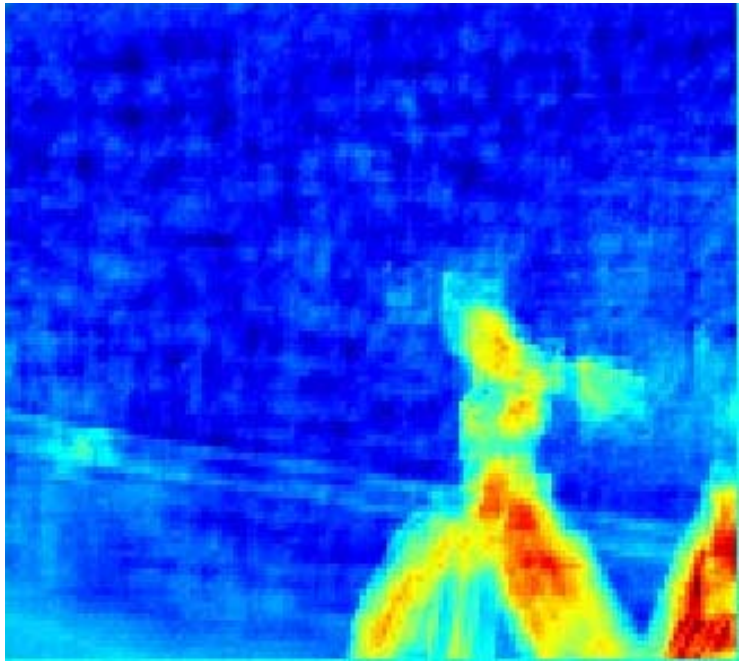
	MNIST	20 newsgroups	Isolet
k q-metrics (m)	$5 \cdot 300$	$5 \cdot 820$	$5 \cdot 23$
k q-metrics	300	820	23

Timings in seconds, 40 passes through the data

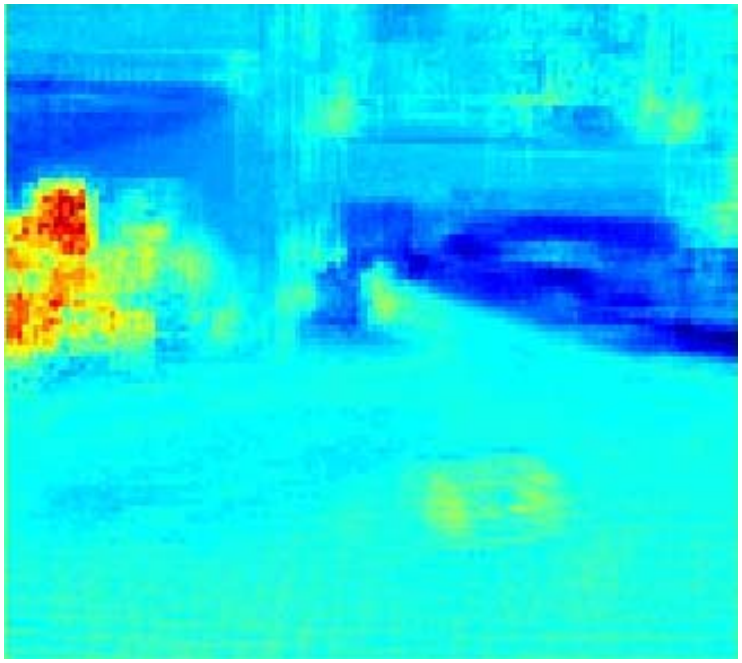
Bikes from the Graz image database:

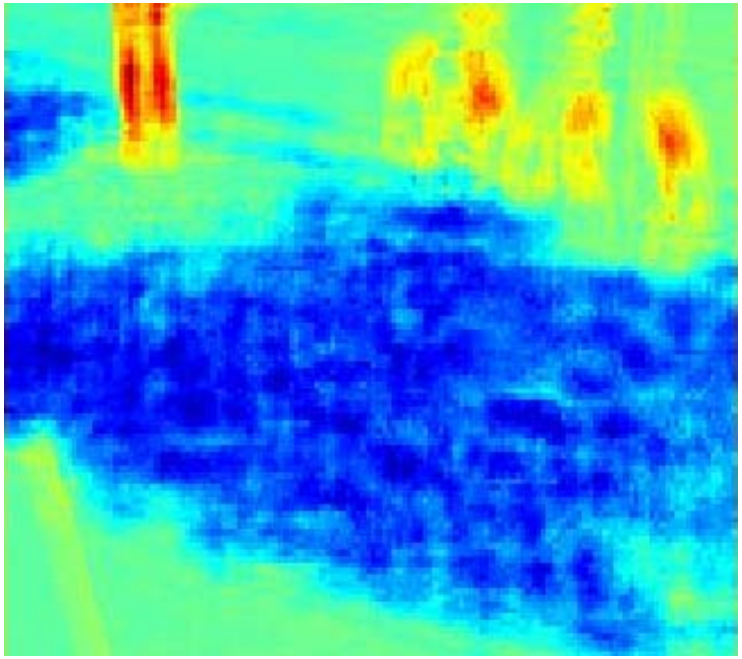


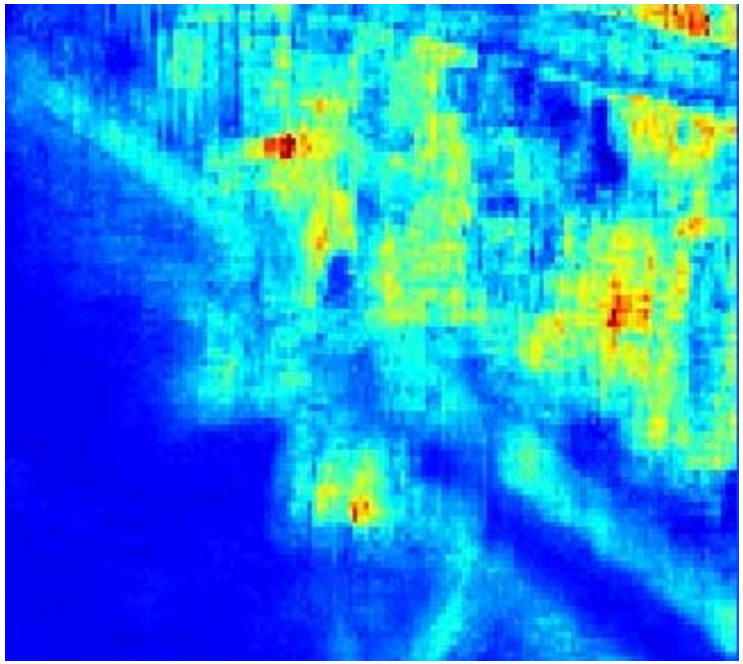
Precision-recall curve for the Graz bikes vs background pixelwise classification. Blue is the proposed method, red dashed is the method in Tuytelaars et al. 2007, and green dashed is the method in Pantofaru et al. 2006.



Note that all decisions are made locally, in a small patch around the pixel in question; also note training data does not have ground truth segmentations.







- Much work to be done.

- Most important problem: a semi-supervised method.