The Log-Determinant Divergence and its Applications

Inderjit S. Dhillon
University of Texas at Austin

Householder Symposium XVII
Zeuthen, Germany
June 3, 2008

Joint work with Jason Davis, Kristen Grauman, Prateek Jain, Brian Kulis, Suvrit Sra and Joel Tropp

Motivating Application: Image Search

- Thousands to millions of pixels in an image
- 3,000-30,000 human recognizable object categories
- Billions of images indexed by Google Image Search
- How can images be represented?
  - Global representations: One vector per image — pixel intensities, color histograms, etc.
  - Local representations: Detect distinctive interest points and extract descriptors invariant to scale, translation, rotation, illumination, etc.

Image Search: Pyramid Match Kernel

- Use Local Representations to represent each image as a set of fixed-dimensional vectors
- Used the pyramid match kernel of [Grauman and Darrell, 2007] to compute approximate matching between two images
  - Place multidimensional, multi-resolution grid over point sets
  - Compute intersection of multi-dimensional histograms at multiple resolutions
  - Compute match between image u and v as:

$$K(u, v) = \sum_{i=0}^{L} w_i N_i(u, v), \quad \text{where } w_i > w_{i+1} > 0,$$

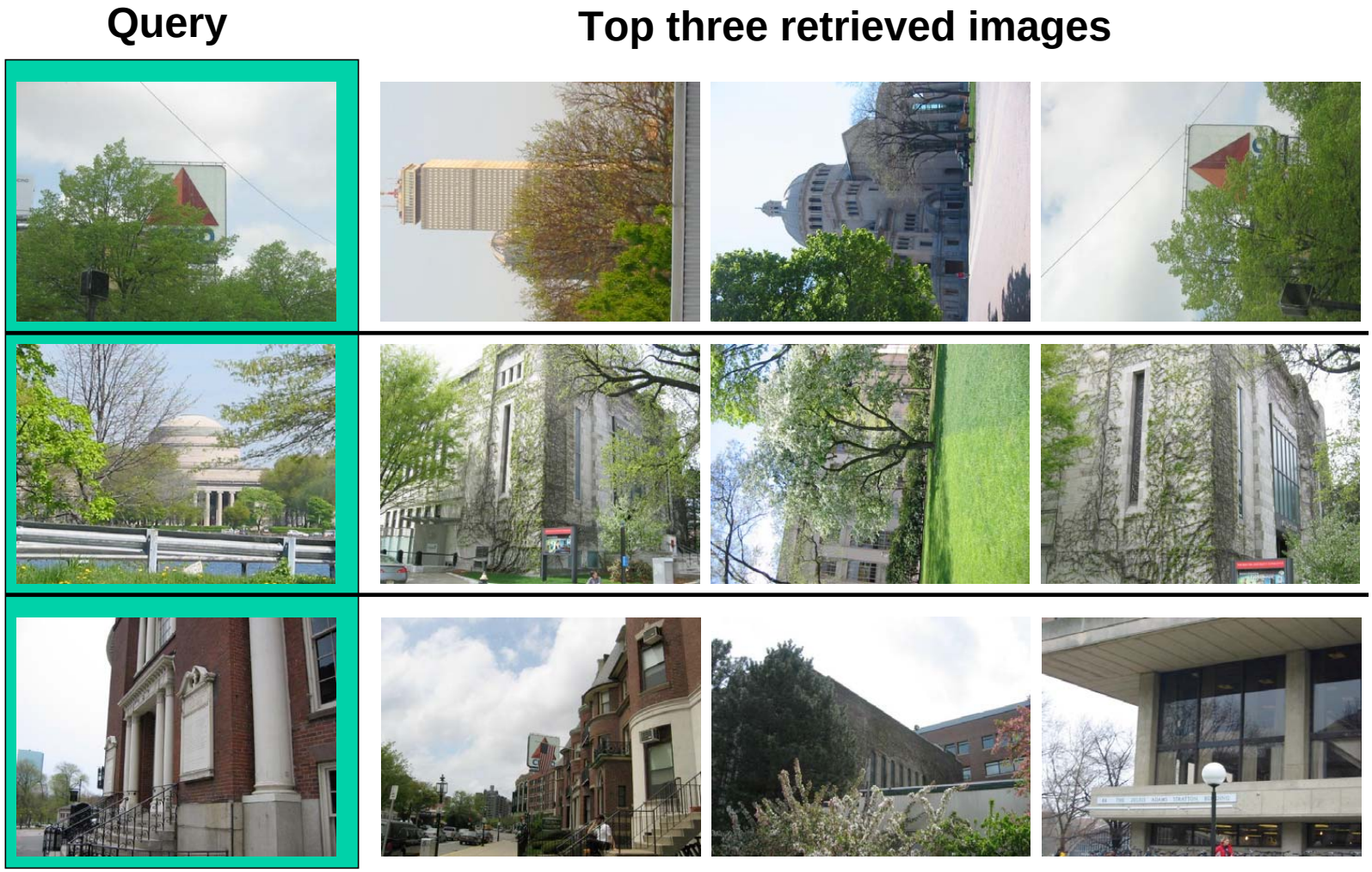
  and N_i is number of newly matched pairs at level i

  - Can show that K is a positive definite matrix (kernel function)

Example query results

Query | **Top three retrieved images**

Example (bad) query results



Example: Image Search Is Not Perfect...

Example: Image Search Is Not Perfect...

charlie van loan - Google Bilder

Web **Bilder** Maps News Shopping Mail Mehr ▼

Anmelden

Google

Erweiterte Bildsuche
Einstellungen

Bilder Ergebnisse **1** - **20** von ungefähr **163.000** für **charlie van loan**. (**0,05** Sekunden)

Anzeigen: **Alle Größen** - Extra groß - groß - mittel - klein

Charles Van Loan
450 x 600 - 241k - gif
www.cs.cornell.edu

Charles Van Loan
Professor
256 x 256 - 50k - gif
www.cs.cornell.edu
[Mehr von www.cs.cornell.edu]

Top to bottom: Joe **Van Loan**, **...**
438 x 594 - 19k - jpg
inkspots.ca

... guitar; Joe **Van Loan**, 1st tenor; **...**
247 x 249 - 7k - jpg
inkspots.ca

... Van Dooren, **Charlie Van Loan**, **...**
489 x 312 - 27k - jpg
www.cs.umd.edu

Van Loan and Acord
216 x 388 - 10k - jpg
gaslight.mtroyal.ca

by **Charles** E. **Van Loan**
400 x 400 - 43k - jpg
gaslight.mtroyal.ca

... Charles Van Loan, Dennis West, **...**
308 x 250 - 29k - jpg
www.northerninitiatives.com

... Charles F. **Van Loan**
501 x 798 - 62k - jpg

Gene H. Golub and **Charles** F. **Van ...**

Author: **Charles** E. **Van Loan**
405 x 500 - 19k - jpg

Patricia **Van Loan**. Manotick ? - 2007

Improving the Kernel

- **The Matrix Nearness Problem:**

$$\begin{aligned}
&\min\ \text{loss}(K, K_0) \\
&(\mathbf{e}_i - \mathbf{e}_j)^T K (\mathbf{e}_i - \mathbf{e}_j) \leq u \quad \text{if } (i,j) \in S \text{ [similarity constraints]} \\
&(\mathbf{e}_i - \mathbf{e}_j)^T K (\mathbf{e}_i - \mathbf{e}_j) \geq \ell \quad \text{if } (i,j) \in D \text{ [dissimilarity constraints]} \\
&K \succeq 0
\end{aligned}$$

- Learn kernel matrix K that is "close" to the baseline kernel matrix K_0
- Other linear constraints on K are possible
- Constraints can arise from various scenarios
  - Unsupervised: Click-through feedback
  - Semi-supervised: must-link and cannot-link constraints
  - Supervised: points in the same class have "small" distance, etc.

Improving the Kernel

- **The Matrix Nearness Problem:**

$$\begin{aligned}
&\min\ \text{loss}(K, K_0) \\
&(\mathbf{e}_i - \mathbf{e}_j)^T K (\mathbf{e}_i - \mathbf{e}_j) \leq u \quad \text{if } (i,j) \in S \text{ [similarity constraints]} \\
&(\mathbf{e}_i - \mathbf{e}_j)^T K (\mathbf{e}_i - \mathbf{e}_j) \geq \ell \quad \text{if } (i,j) \in D \text{ [dissimilarity constraints]} \\
&K \succeq 0
\end{aligned}$$

- Learn kernel matrix K that is "close" to the baseline kernel matrix K_0
- Other linear constraints on K are possible
- Constraints can arise from various scenarios
  - Unsupervised: Click-through feedback
  - Semi-supervised: must-link and cannot-link constraints
  - Supervised: points in the same class have "small" distance, etc.
- QUESTION: What should "loss" be?

Improving the Kernel

- **The Matrix Nearness Problem:**

$$\begin{aligned}
&\min\ \text{loss}(K, K_0) \\
&(\mathbf{e}_i - \mathbf{e}_j)^T K (\mathbf{e}_i - \mathbf{e}_j) \leq u \quad \text{if } (i,j) \in S \text{ [similarity constraints]} \\
&(\mathbf{e}_i - \mathbf{e}_j)^T K (\mathbf{e}_i - \mathbf{e}_j) \geq \ell \quad \text{if } (i,j) \in D \text{ [dissimilarity constraints]} \\
&K \succeq 0
\end{aligned}$$

- Learn kernel matrix K that is "close" to the baseline kernel matrix K_0
- Other linear constraints on K are possible
- Constraints can arise from various scenarios
  - Unsupervised: Click-through feedback
  - Semi-supervised: must-link and cannot-link constraints
  - Supervised: points in the same class have "small" distance, etc.
- QUESTION: What should "loss" be?
- We use "loss" to be the LogDet Divergence

What is the LogDet Divergence?

Distance between Positive Definite Matrices

- Frobenius Distance:

$$D_{Frob} = \|X - Y\|_F$$

Distance between Positive Definite Matrices

- Frobenius Distance:

$$D_{Frob} = \|X - Y\|_F$$

- LogDet Divergence:

$$D_{\ell d}(X, Y) = \text{trace}(XY^{-1}) - \log\det(XY^{-1}) - d,$$

Distance between Positive Definite Matrices

- Frobenius Distance:

$$D_{Frob} = \|X - Y\|_F$$

- LogDet Divergence:

$$\begin{aligned} D_{\ell d}(X, Y) &= \text{trace}(XY^{-1}) - \log\det(XY^{-1}) - d, \\ &= \text{trace}(Y^{-1/2}XY^{-1/2}) - \log\det(Y^{-1/2}XY^{-1/2}) - d, \end{aligned}$$

Distance between Positive Definite Matrices

- Frobenius Distance:

$$D_{Frob} = \|X - Y\|_F$$

- LogDet Divergence:

$$\begin{aligned} D_{\ell d}(X, Y) &= \text{trace}(XY^{-1}) - \log\det(XY^{-1}) - d, \\ &= \text{trace}(Y^{-1/2}XY^{-1/2}) - \log\det(Y^{-1/2}XY^{-1/2}) - d, \\ &= \text{trace}(Y^{-1/2}XY^{-1/2} - \log Y^{-1/2}XY^{-1/2} - I), \end{aligned}$$

Distance between Positive Definite Matrices

- Frobenius Distance:

$$D_{Frob} = \|X - Y\|_F$$

- LogDet Divergence:

$$\begin{aligned} D_{\ell d}(X, Y) &= \text{trace}(XY^{-1}) - \log\det(XY^{-1}) - d, \\ &= \text{trace}(Y^{-1/2}XY^{-1/2}) - \log\det(Y^{-1/2}XY^{-1/2}) - d, \\ &= \text{trace}(Y^{-1/2}XY^{-1/2} - \log Y^{-1/2}XY^{-1/2} - I), \\ &= \sum_i \sum_j (\mathbf{v}_i^T \mathbf{u}_j)^2 \left(\frac{\lambda_i}{\theta_j} - \log \frac{\lambda_i}{\theta_j} - 1 \right), \end{aligned}$$

where $X = V\Lambda V^T$ and $Y = U\Theta U^T$

LogDet Divergence: Basic Properties

$$D_{\ell d}(X, Y) \quad = \quad \text{trace}(Y^{-1/2} X Y^{-1/2}) - \log\det(Y^{-1/2} X Y^{-1/2}) - d,$$

- Can be used as a measure of distance
  - Positive, and zero iff $X = Y$
  - But not symmetric, and triangle inequality does not hold
- Convex in first argument (not in second)
- Pythagorean Property holds: Given Y and a convex set Ω,

$$D_{\ell d}(X, Y) \geq D_{\ell d}(X, P_\Omega(Y)) + D_{\ell d}(P_\Omega(Y), Y), \quad \text{holds for all } X \in \Omega$$

- Definition can be extended to semi-definite matrices

LogDet Divergence: Scale Invariance

$$D_{\ell d}(X, Y) \quad = \quad \text{trace}(Y^{-1/2} X Y^{-1/2}) - \log\det(Y^{-1/2} X Y^{-1/2}) - d,$$

- Scale-invariance

$$D_{\ell d}(X, Y) = D_{\ell d}(\alpha X, \alpha Y), \quad \alpha \geq 0$$

- In fact, for any invertible M

$$D_{\ell d}(X, Y) = D_{\ell d}(M^T X M, M^T Y M)$$

- In particular,

$$D_{\ell d}(X, Y) = D_{\ell d}(Y^{-1/2} X Y^{-1/2}, I)$$

LogDet and the Condition Number

$$\begin{aligned} D_{\ell d}(X, I) &= \operatorname{trace}(X) - \log\det(X) - d, \\ &= \sum_{i=1}^{d} \left(\lambda_i - \log \lambda_i - 1 \right) \end{aligned}$$

LogDet and the Condition Number

$$\begin{aligned} D_{\ell d}(X, I) &= \operatorname{trace}(X) - \log\det(X) - d, \\ &= \sum_{i=1}^{d} \left(\lambda_i - \log \lambda_i - 1 \right) \end{aligned}$$

- Now, $x - \log x \geq 1$ with equality at $x = 1$

LogDet and the Condition Number

$$\begin{aligned} D_{\ell d}(X, I) &= \operatorname{trace}(X) - \log\det(X) - d, \\ &= \sum_{i=1}^{d} \left(\lambda_i - \log \lambda_i - 1 \right) \end{aligned}$$

- Now, $x - \log x \geq 1$ with equality at $x = 1$
- Also, $x - \log x \geq \log x + 1 - \log 4$ with equality at $x = 2$

LogDet and the Condition Number

$$\begin{aligned} D_{\ell d}(X, I) &= \operatorname{trace}(X) - \log\det(X) - d, \\ &= \sum_{i=1}^{d} \left(\lambda_i - \log \lambda_i - 1 \right) \end{aligned}$$

- Now, $x - \log x \geq 1$ with equality at $x = 1$
- Also, $x - \log x \geq \log x + 1 - \log 4$ with equality at $x = 2$
- Letting, $\lambda_1 \geq \lambda_2 \cdots \geq \lambda_d > 0$

$$\begin{aligned} D_{\ell d}(X, I) &\geq (\log \lambda_1 + 1 - \log 4) - (\log \lambda_d + 1), \\ \implies \quad cond(X) &\leq 4 \exp D_{\ell d}(X, I) \end{aligned}$$

- Thus, LogDet yields an upper bound on the condition number

Where does LogDet occur?

The Gaussian Connection — Maximum Likelihood

- Suppose $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_m$ are drawn from:

$$p(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{d/2}(\det\boldsymbol{\Sigma})^{1/2}} \exp\left\{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right\}$$

The Gaussian Connection — Maximum Likelihood

- Suppose $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_m$ are drawn from:

$$p(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{d/2}(\det\boldsymbol{\Sigma})^{1/2}} \exp\left\{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right\}$$

- Log-Likelihood:

$$\begin{aligned} L(\boldsymbol{\mu}, \boldsymbol{\Sigma}) &= \prod_{i=1}^{m} p(\mathbf{x}_i|\boldsymbol{\mu}, \boldsymbol{\Sigma}) \\ &\propto \exp\left\{-\frac{m}{2}\left(D_{\ell d}(\boldsymbol{\Sigma}^{-1}, \bar{\mathbf{S}}^{-1}) + (\bar{\boldsymbol{\mu}}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\bar{\boldsymbol{\mu}}-\boldsymbol{\mu})\right)\right\}, \end{aligned}$$

where $\bar{\boldsymbol{\mu}} = \frac{1}{m}\sum_{i=1}^{m}\mathbf{x}_i$ and $\bar{\mathbf{S}} = \frac{1}{m}\sum_{i=1}^{m}(\mathbf{x}_i-\bar{\boldsymbol{\mu}})(\mathbf{x}_i-\bar{\boldsymbol{\mu}})^T$

The Gaussian Connection — Maximum Likelihood

- Suppose $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_m$ are drawn from:

$$p(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{d/2}(\det\boldsymbol{\Sigma})^{1/2}} \exp\left\{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right\}$$

- Log-Likelihood:

$$\begin{aligned} L(\boldsymbol{\mu}, \boldsymbol{\Sigma}) &= \prod_{i=1}^{m} p(\mathbf{x}_i|\boldsymbol{\mu}, \boldsymbol{\Sigma}) \\ &\propto \exp\left\{-\frac{m}{2}\left(D_{\ell d}(\boldsymbol{\Sigma}^{-1}, \bar{\mathbf{S}}^{-1}) + (\bar{\boldsymbol{\mu}}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\bar{\boldsymbol{\mu}}-\boldsymbol{\mu})\right)\right\}, \end{aligned}$$

  where $\bar{\boldsymbol{\mu}} = \frac{1}{m}\sum_{i=1}^{m}\mathbf{x}_i$ and $\bar{\mathbf{S}} = \frac{1}{m}\sum_{i=1}^{m}(\mathbf{x}_i-\bar{\boldsymbol{\mu}})(\mathbf{x}_i-\bar{\boldsymbol{\mu}})^T$

- Thus, $\bar{\boldsymbol{\mu}}$ and $\bar{\mathbf{S}}$ are the maximum likelihood estimates of the mean & covariance matrix, respectively

Further Connections in Statistics

- Wishart Distribution — Given m samples from a Gaussian distribution, the pdf of the sample covariance matrix may be written as:

$$p(\mathbf{S}, m) \propto \exp\left\{-\frac{m}{2} D_{\ell d}(\boldsymbol{\Sigma}^{-1}, \mathbf{S}^{-1}) - \frac{d+1}{2} \log\det\mathbf{S}\right\}$$

- Differential Relative Entropy between two Multivariate Gaussians:

$$\int p(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}_0) \log\left(\frac{p(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}_0)}{p(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma})}\right) d\mathbf{x} = \frac{1}{2} D_{\ell d}(\boldsymbol{\Sigma}, \boldsymbol{\Sigma}_0)$$

- [James and Stein, 1961] LogDet divergence is known as Stein's loss in the statistics community

Quasi-Newton Optimization

- LogDet Divergence arises in the BFGS and DFP updates
  - Quasi-Newton methods
  - Approximate Hessian of the function to be minimized
- [Fletcher, 1991] BFGS update can be shown to optimize:

$$\begin{aligned} \min_{B} \quad & D_{\ell d}(B, B_t) \\ \text{subject to} \quad & B\, s_t = y_t \quad (\text{“Secant Equation”}) \end{aligned}$$

- $s_t = x_{t+1} - x_t$, $y_t = \nabla f_{t+1} - \nabla f_t$
- Closed-form solution:

$$B_{t+1} = B_t - \frac{B_t s_s s_t^T B_t}{s_t^T B_t s_t} + \frac{y_t y_t^T}{s_t^T y_t}$$

- Similar form for DFP update

How do we use LogDet?

Kernel Learning

- **The Matrix Nearness Problem:**

$$
\begin{aligned}
&\min\ D_{\ell d}(K, K_0) \\
&(\mathbf{e}_i - \mathbf{e}_j)^T K (\mathbf{e}_i - \mathbf{e}_j) \leq u \quad \text{if } (i,j) \in S \text{ [similarity constraints]} \\
&(\mathbf{e}_i - \mathbf{e}_j)^T K (\mathbf{e}_i - \mathbf{e}_j) \geq \ell \quad \text{if } (i,j) \in D \text{ [dissimilarity constraints]} \\
&K \succeq 0
\end{aligned}
$$

Successive Projection-Correction Algorithm

- Algorithm: project successively onto each linear constraint followed by correction — **c**onverges to globally optimal solution
- Each projection updates the Kernel matrix:

$$\begin{aligned} \min_{K} \quad & D_{\ell d}(K, K_t) \\ \text{s.t.} \quad & (\mathbf{e}_i - \mathbf{e}_j)^T K (\mathbf{e}_i - \mathbf{e}_j) \leq u \end{aligned}$$

- Can be solved by rank-one update:

$$\begin{aligned} K_{t+1} &= K_t + \theta_t K_t (\mathbf{e}_i - \mathbf{e}_j)(\mathbf{e}_i - \mathbf{e}_j)^T K_t \\ \Longrightarrow \quad K^* &= K_0 + K_0 \Theta K_0 \end{aligned}$$

- Advantages:
  - Automatic enforcement of positive semidefiniteness
  - Simple, closed-form projections (θ_t computable in closed form)
  - No eigenvalue/eigenvector calculation
  - Easy to incorporate slack for each constraint

Enhancements

- Impose structure on K:
  - low-rank
  - K_0 + low-rank
  - Each iteration costs only $O(r^2)$, where r is rank
- Extension to new data points:
  - Learned kernel can be shown to be of the form

$$K(\mathbf{x}, \mathbf{y}) = K_0(\mathbf{x}, \mathbf{y}) + \sum_i \sum_j \theta_{ij} K_0(\mathbf{x}, \mathbf{x}_i) K_0(\mathbf{y}, \mathbf{x}_j)$$

- Fast Nearest-Neighbor Search
  - Can incorporate "locality-sensitive hashing"
- Online algorithms
  - Constraints are presented incrementally

Results

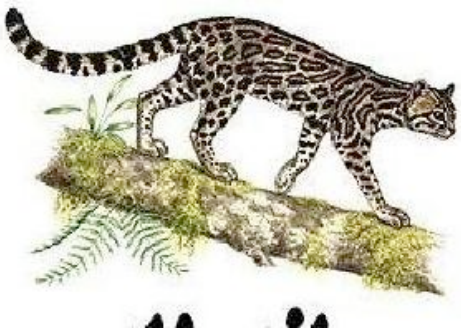
Application 1: Image Recognition

Data Set: Caltech 101

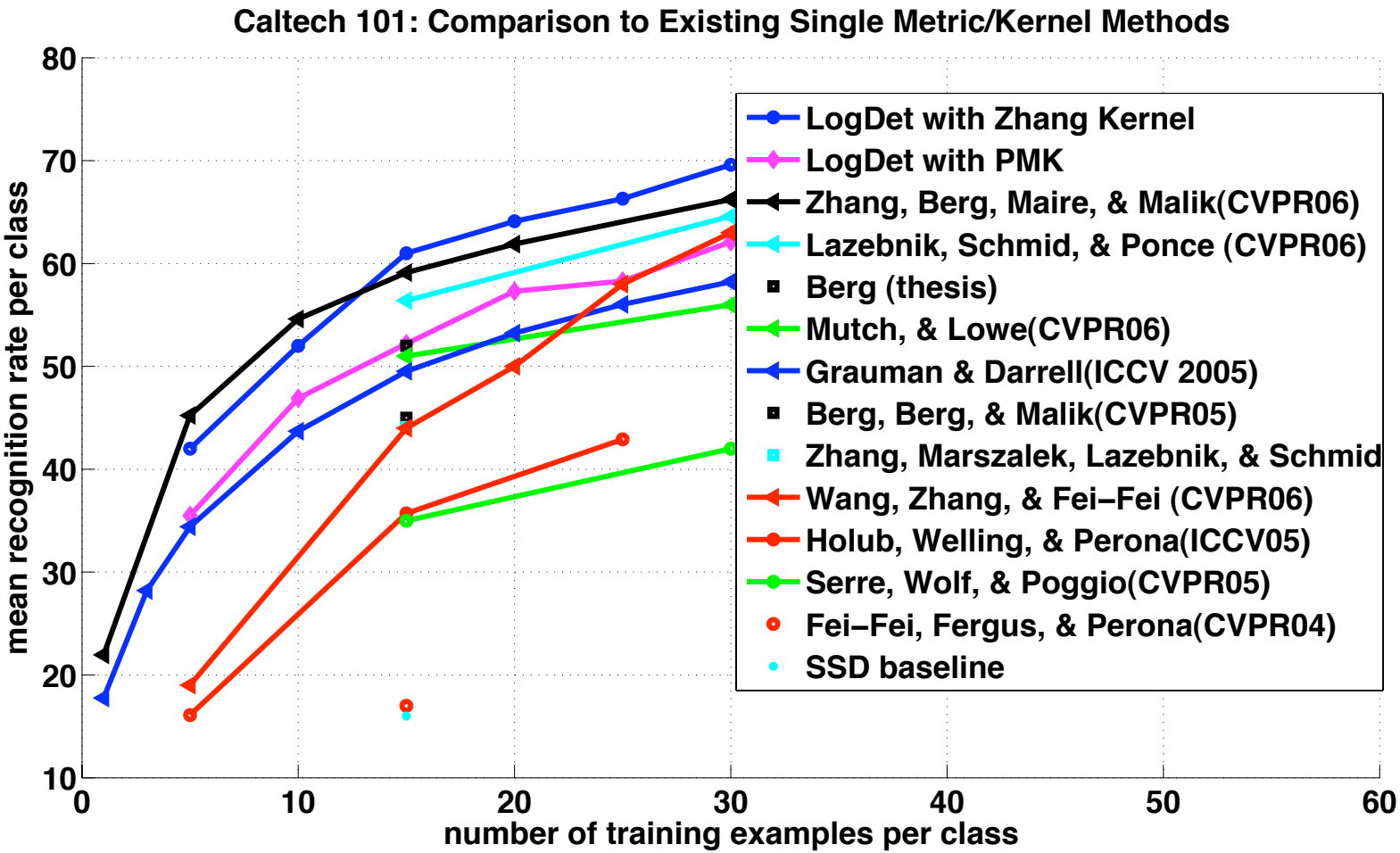
- Standard benchmark for multi-category image recognition
- 101 classes of images
- Wide variance in pose etc.
- Challenging data set

Experimental Setup

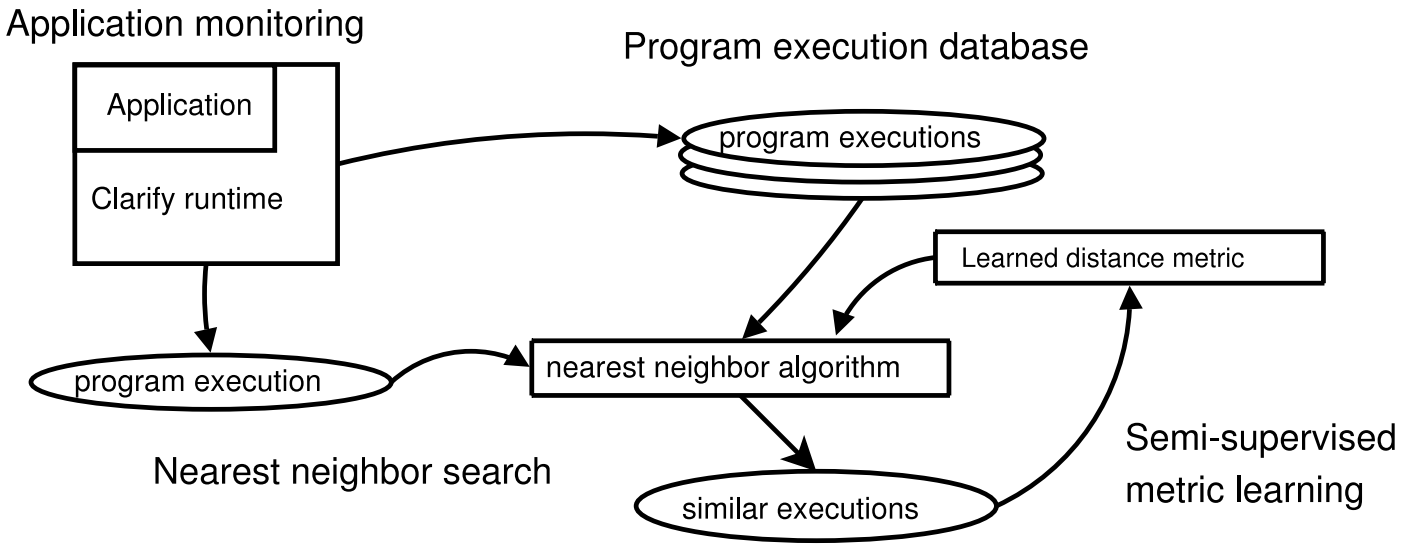
- 5, 10, 15 & 30 images per class in training set; rest in test set
- Performed 1-NN using original kernels and learned kernels



Results: Image Recognition



Application 2: Automating Software Support

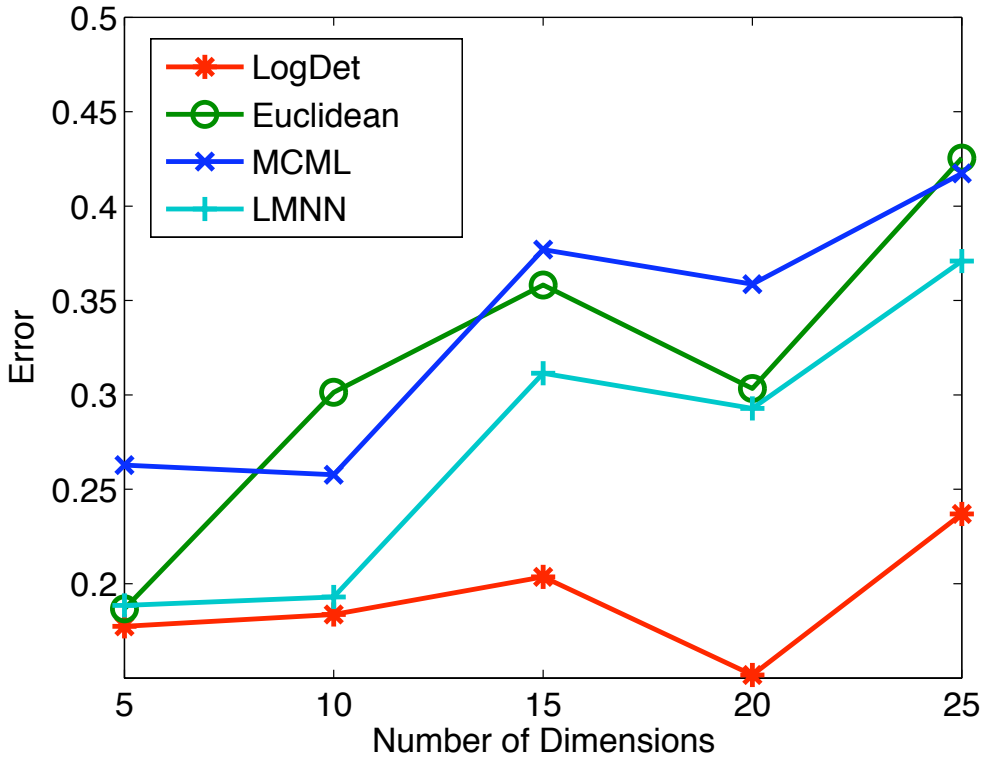


- Clarify: system to compare a user's program execution to existing executions in a database to automate software support
- Need appropriate notion of "distance" between program executions

Results: Clarify

- Representation: System collects program features during run-time
  - Function counts
  - Call-site counts
  - Counts of program paths
  - Program execution represented as a vector of counts
- Class labels: Program execution errors
- Nearest neighbor software support
  - Match program executions
  - Underlying distance measure should reflect this similarity

LaTeX Results



Conclusions

LogDet Divergence

- Previously used in Statistics & Optimization
- Has intriguing properties

Applications

- Leads to new matrix nearness problems
- Successive projection-correction algorithm
- Each projection can be computed efficiently

Challenges

- Faster solutions to matrix nearness problem – interior point methods?
- Apply to find low-rank correlation matrices

References

"Learning Low-Rank Kernel Matrices", B. Kulis, M. A. Sustik, and I. S. Dhillon. International Conference on Machine Learning (ICML), pages 505-512, July 2006 (longer version submitted to journal).

"Information-Theoretic Metric Learning", J. Davis, B. Kulis, P. Jain, S. Sra, and I. S. Dhillon. International Conference on Machine Learning (ICML), pages 209-216, June 2007.

"Structured Metric Learning for High-Dimensional Problems", J. Davis and I. S. Dhillon. ACM International Conference on Knowledge Discovery and Data Mining(KDD), August 2008.

"Fast Image Search for Learned Metrics", P. Jain, B. Kulis, and K. Grauman. IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 2008.

"The Pyramid Match Kernel: Efficient Learning with Sets of Features", K. Grauman and T. Darrell. Journal of Machine Learning Research (JMLR), vol. 8, pages 725–760, Apr 2007.

"Matrix Nearness Problems using Bregman Divergences", I. S. Dhillon and J. A. Tropp. SIAM Journal on Matrix Analysis and its Applications, vol. 29, no. 4, pages 1120-1146, Dec 2007.