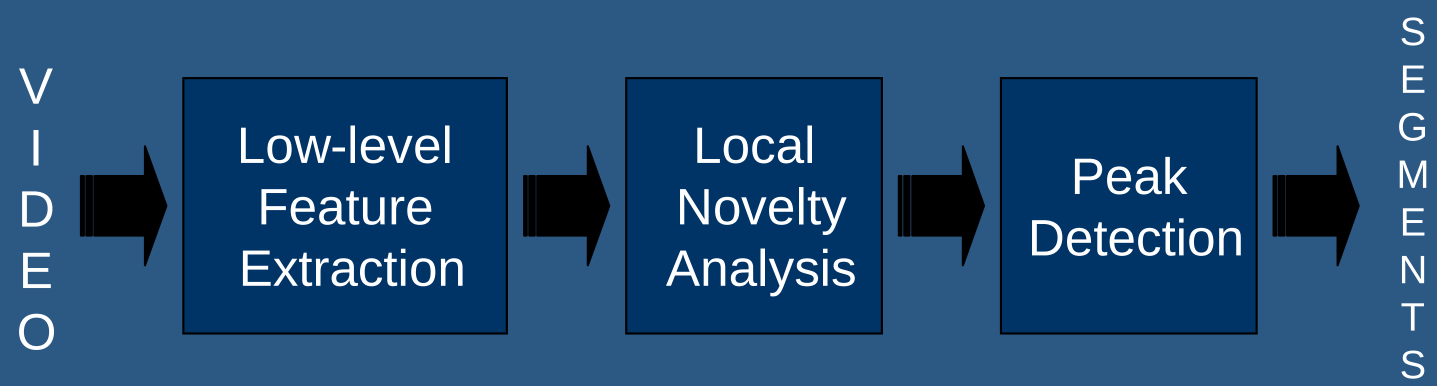
shot boundary detection combining similarity analysis and classification

Matthew Cooper[1], Ting Liu[2], and Eleanor Rieffel[1]

[1]FX Palo Alto Laboratory
http://www.fxpal.com

[2]Dept. of Computer Science
Carnegie Melon University
http://www.autonlab.org

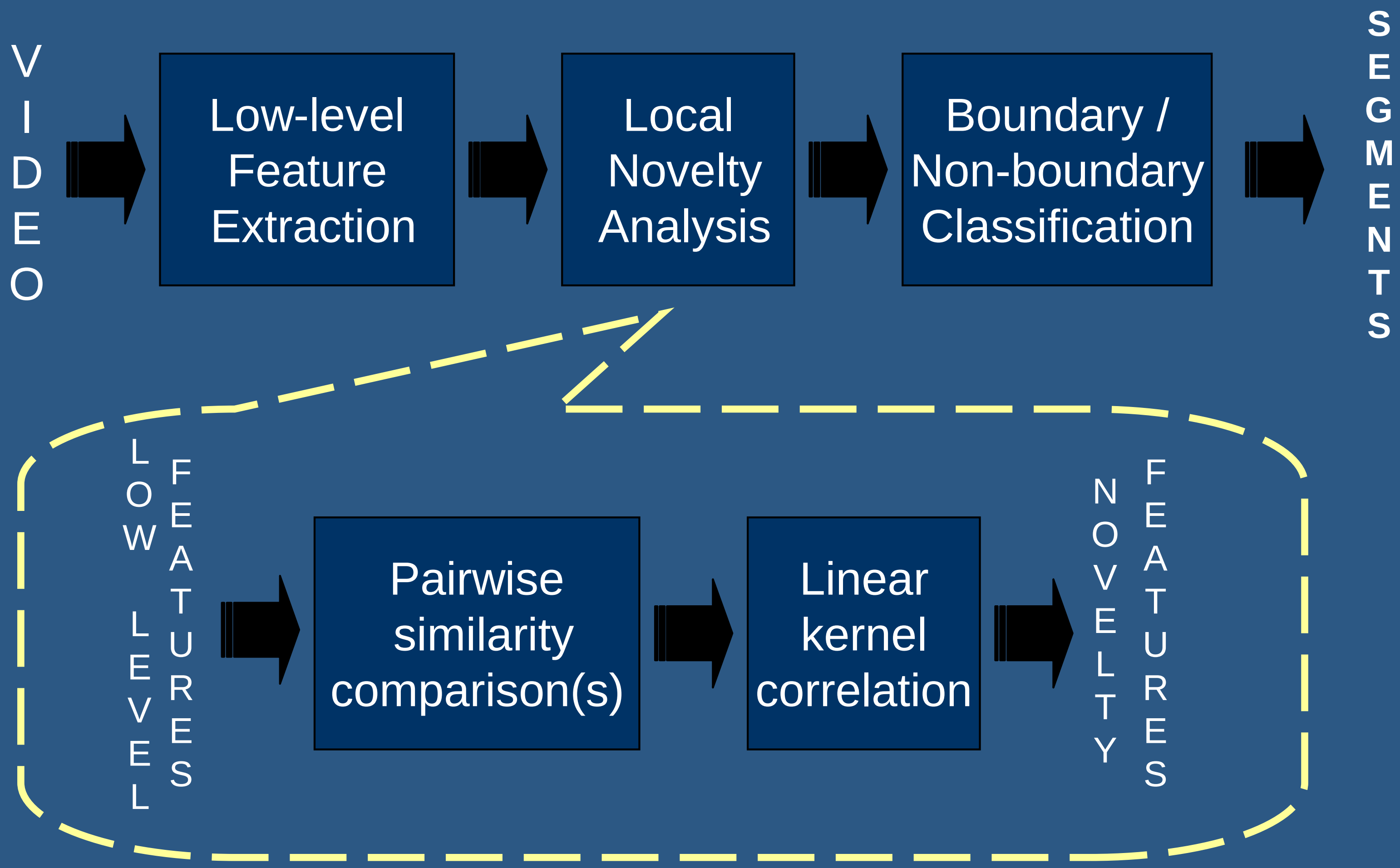
traditional video segmentation



- what’s working and what’s not?
  - features are YUV histograms (block and global)
  - replace ad hoc peak detection with supervised classification as in [Qi, et al., 2003]

Y. Qi, A. Hauptman, T.Liu. Supervised Classification for Video Shot Segmentation. In *Proc. of IEEE International Conference on Multimedia & Expo*, 2003.

reformulating segmentation



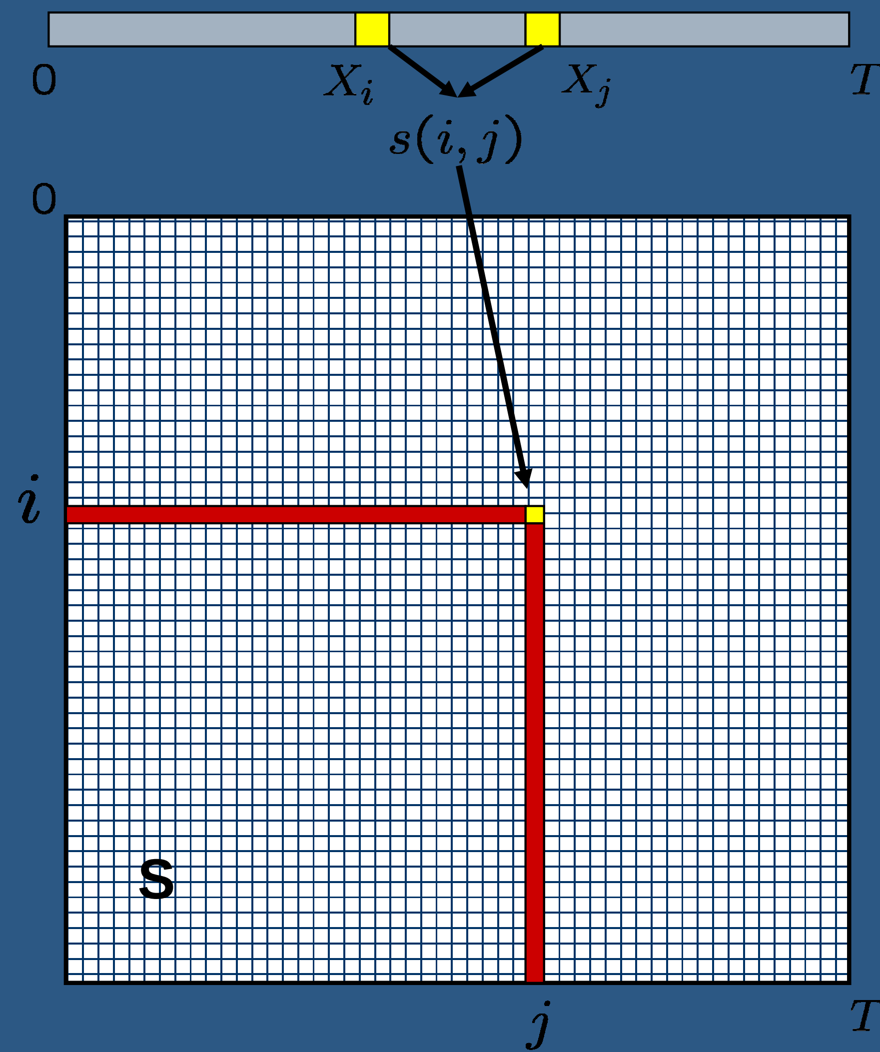
inter-frame similarity analysis

- concatenate YUV histogram features

$$f_i \rightarrow x_i \;\; (x_i \in R^p)$$

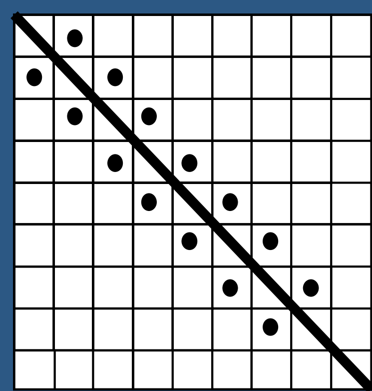
- construct L1 similarity matrix:

$$S(i,j) = \sum_{p=1}^{P} |X_i(p) - X_j(p)|$$



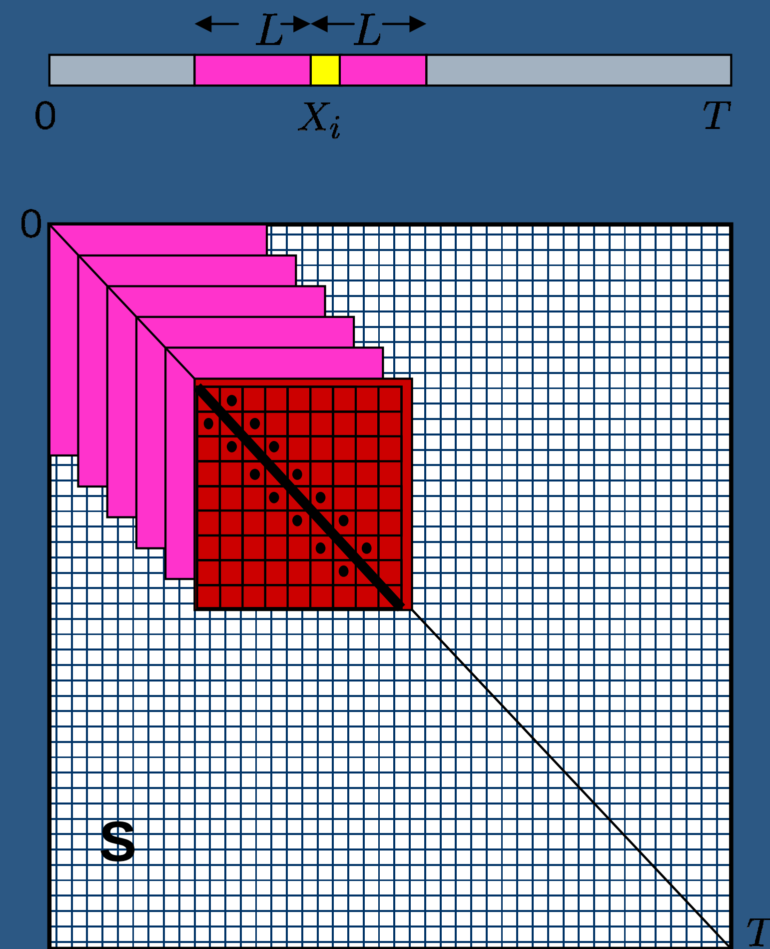
novelty via kernel correlation

- scale-space kernel linearly combines adjacent frame comparisons

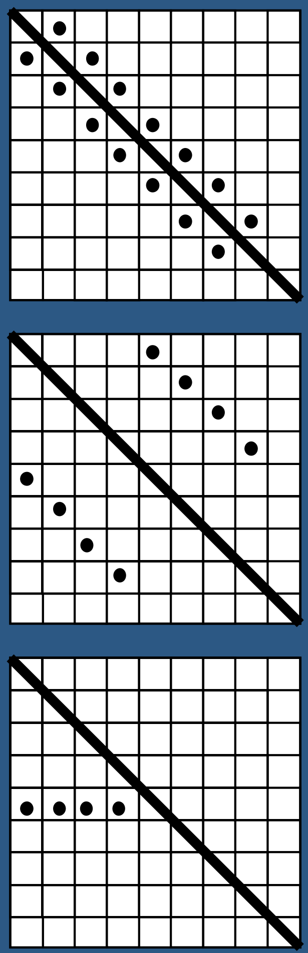


- more generally:

$$\nu(n) = \sum_{l=-L}^{L-1} \sum_{m=-L}^{L-1} \mathbf{K}(l,m)\mathbf{S}(n+l,n+m)$$

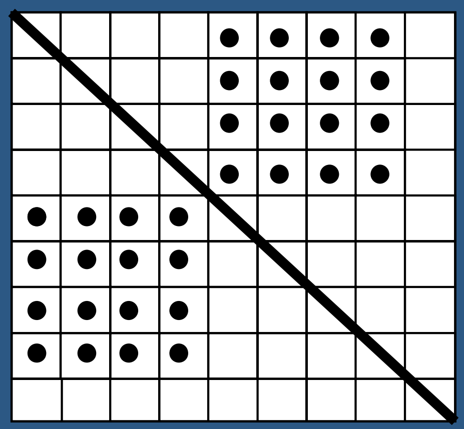


related work: dissimilarity kernels

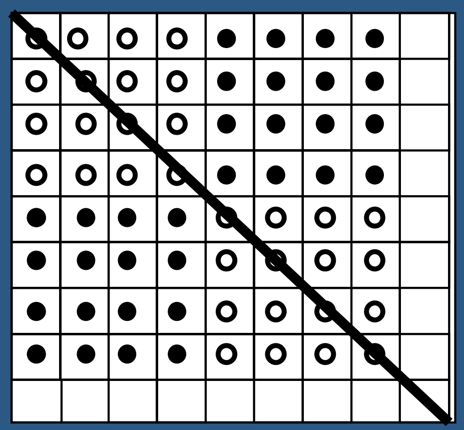


- **s**cale-**s**pace (SS) kernel weights only adjacent inter-frame similarities [e.g. Witkin, 1984]
- **d**iagonal **c**ross-**s**imilarity (DCS) kernel weights inter-frame similarity of pairs *L* frames apart [Pye et al., 1998; Pickering et al., TRECVIDs]
- row (ROW) kernel compares current frame to each frame in local neighborhood [Qi, et al., 2003]

dissimilarity kernels



- cross similarity (CS) kernel is matched filter for ideal dissimilarity boundary



- full similarity (FS) kernel penalizes within-segment dissimilarity [Cooper and Foote, ICIP 2001]

input features for classification

$$\nu(n) = \sum_{l=-L}^{L-1} \sum_{m=-L}^{L-1} \mathbf{K}(l,m)\mathbf{S}(n+l, n+m)$$

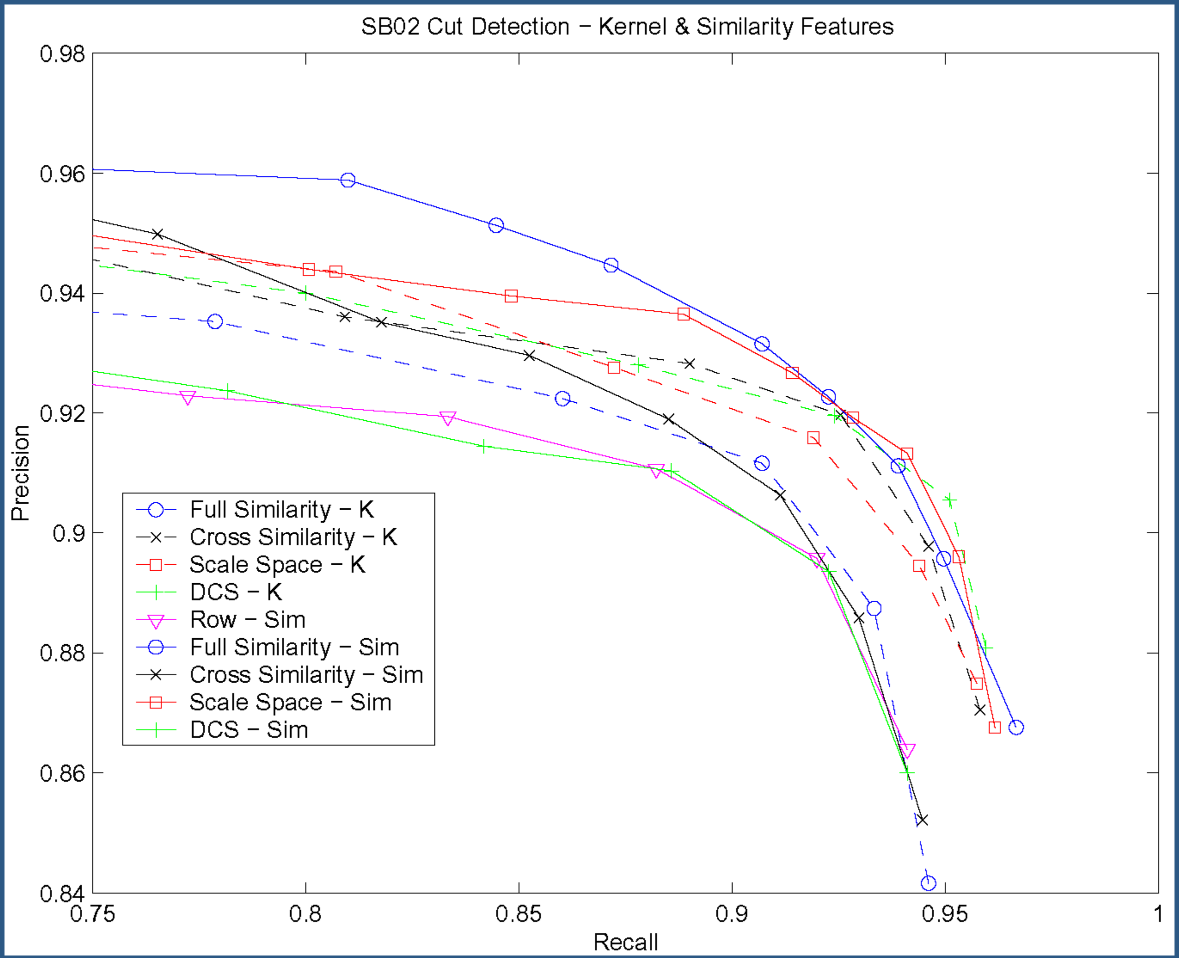
- kernel-based features: concatenate frame-indexed kernel correlations $\nu_L(n)$ for L=2,3,4,5, for both global histogram similarity and block histogram similarity
- raw similarity features: concatenate all raw similarity comparisons that contribute to kernel correlation for L=5 (without linearly combining them)

experimental setup

- efficient exact kNN classifier provided by T. Liu and A. Moore at CMU (http://www.autonlab.org)
  - ball-tree implementation ~ 10 times speedups over naïve kNN
  - for details, see [Liu, Moore, Gray, NIPS 2003]
- TRECVID 2002 test set for **cut** boundary detection
  - almost 6 hours of broadcast news data
  - manual ground truth, 1466 cut boundaries
  - medians from TV02: recall = 0.86, precision = 0.84
  - hold-one-out cross validation, k = 11

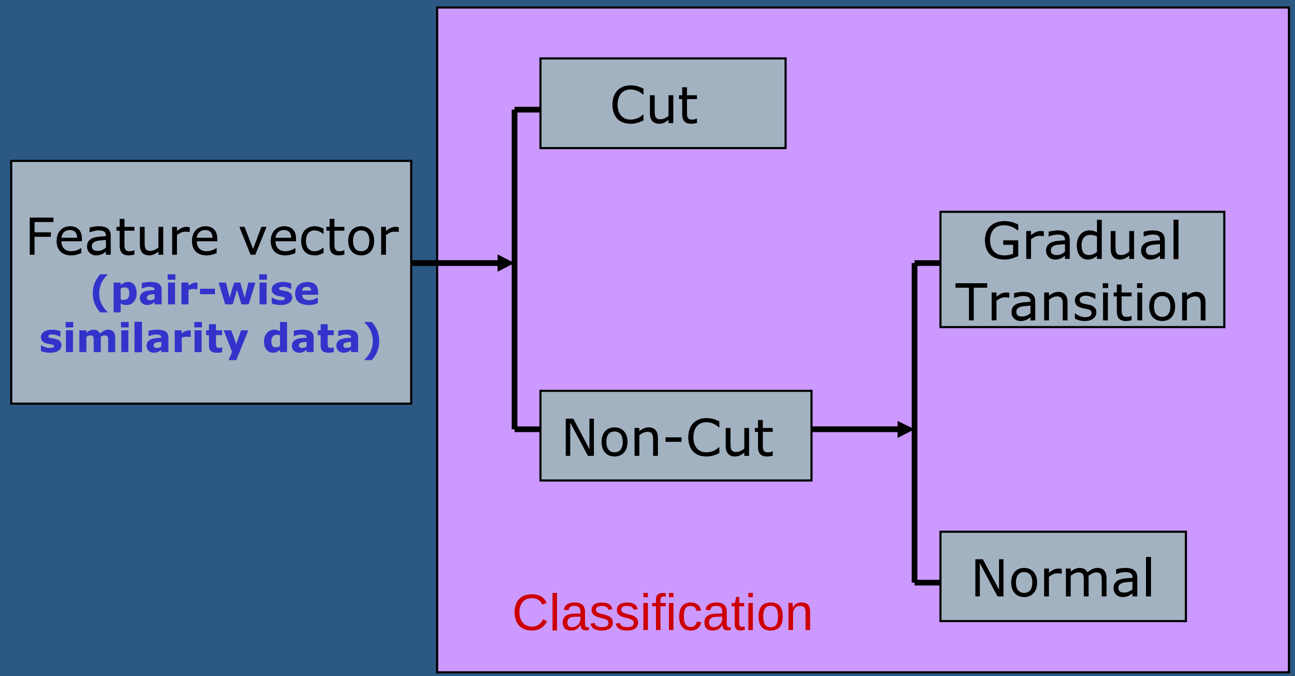
comparative results

- FS similarity features provide most information and achieve best overall performance



setup for SB04

- to extend to cut and gradual detection, we follow two-step binary classification approach in [Qi, et al., 2003]



- unlike prior work no smoothing of classifier outputs, no motion, flash, etc.
- efficient exact kNN classifier k = 11
- 8 CNN and ABC videos from SB03 test set
- hold-one-out cross validation

training – varying the similarity measure

- FS pairwise similarity features used
- 8 ABC and CNN videos in SB03 test set used for training
- testing similarity measures

$$S(i,j) = \sum_{p=1}^{P} |X_i(p) - X_j(p)|$$

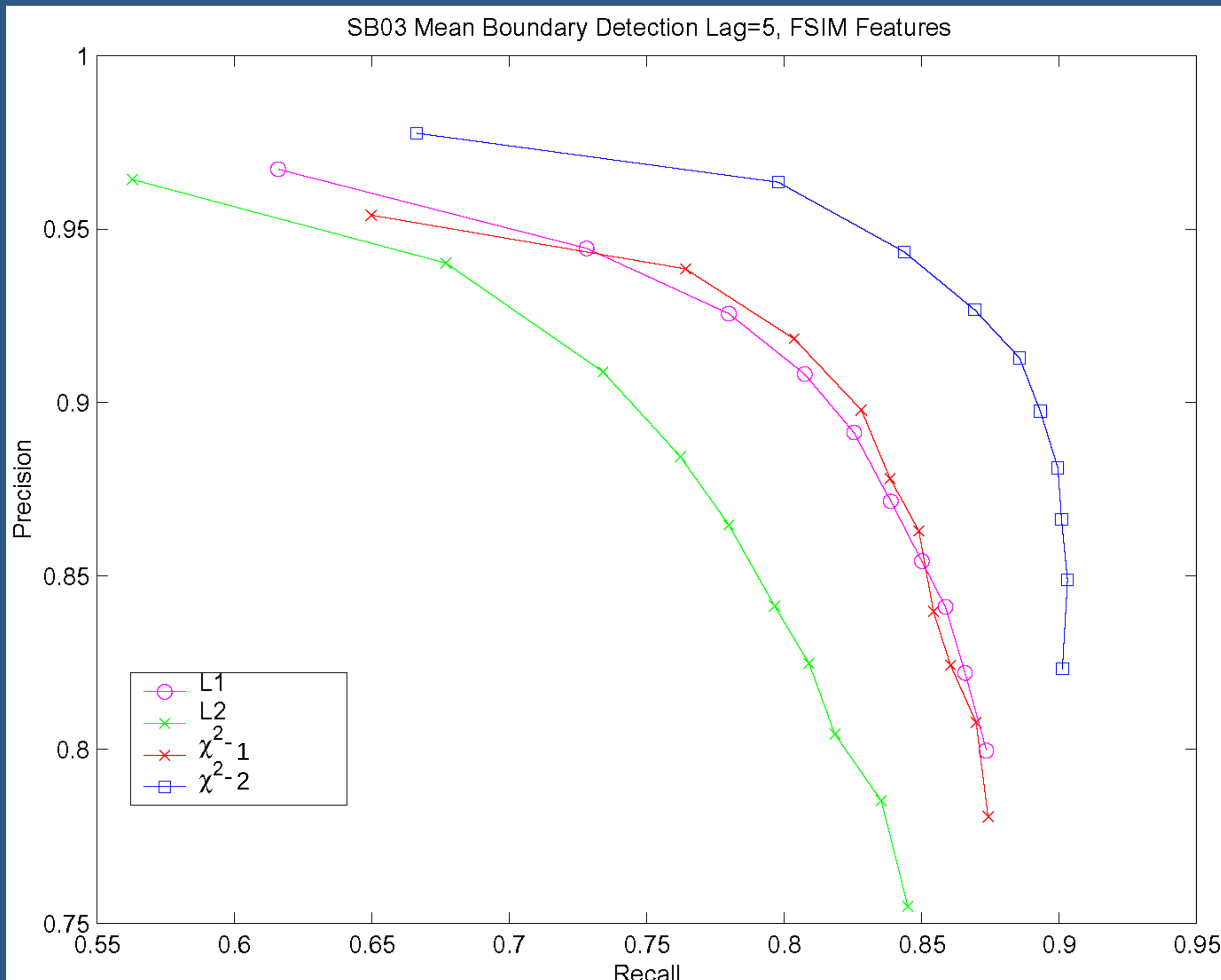
$$S(i,j) = \sqrt{\sum_{p=1}^{P} (X_i(p) - X_j(p))^2}$$

$$S(i,j) = \sum_{p=1}^{P} \frac{(X_i(p) - E_{ij}(p))^2}{(X_i(p) + E_{ij}(p))}$$

$$S(i,j) = \sum_{p=1}^{P} \frac{(X_i(p) - E_{ij}(p))^2}{(X_i(p) + E_{ij}(p))^2}$$

- testing different lag L=5, 10
- random projection for dimension reduction for L=10

comparing similarity measures



training – varying *L*

- L=10 implies FS feature dimensionality is d=380
- problem of fast kNN
  - significant speed-up when d is small: O(1) ~ O(dNlogN)
  - little speed-up when d is large: O(dN²)
- random projection

THM (Johnson-Lindenstrauss lemma) For any $0 < \epsilon < 1$ and any integer N, let d' be a positive integer such that

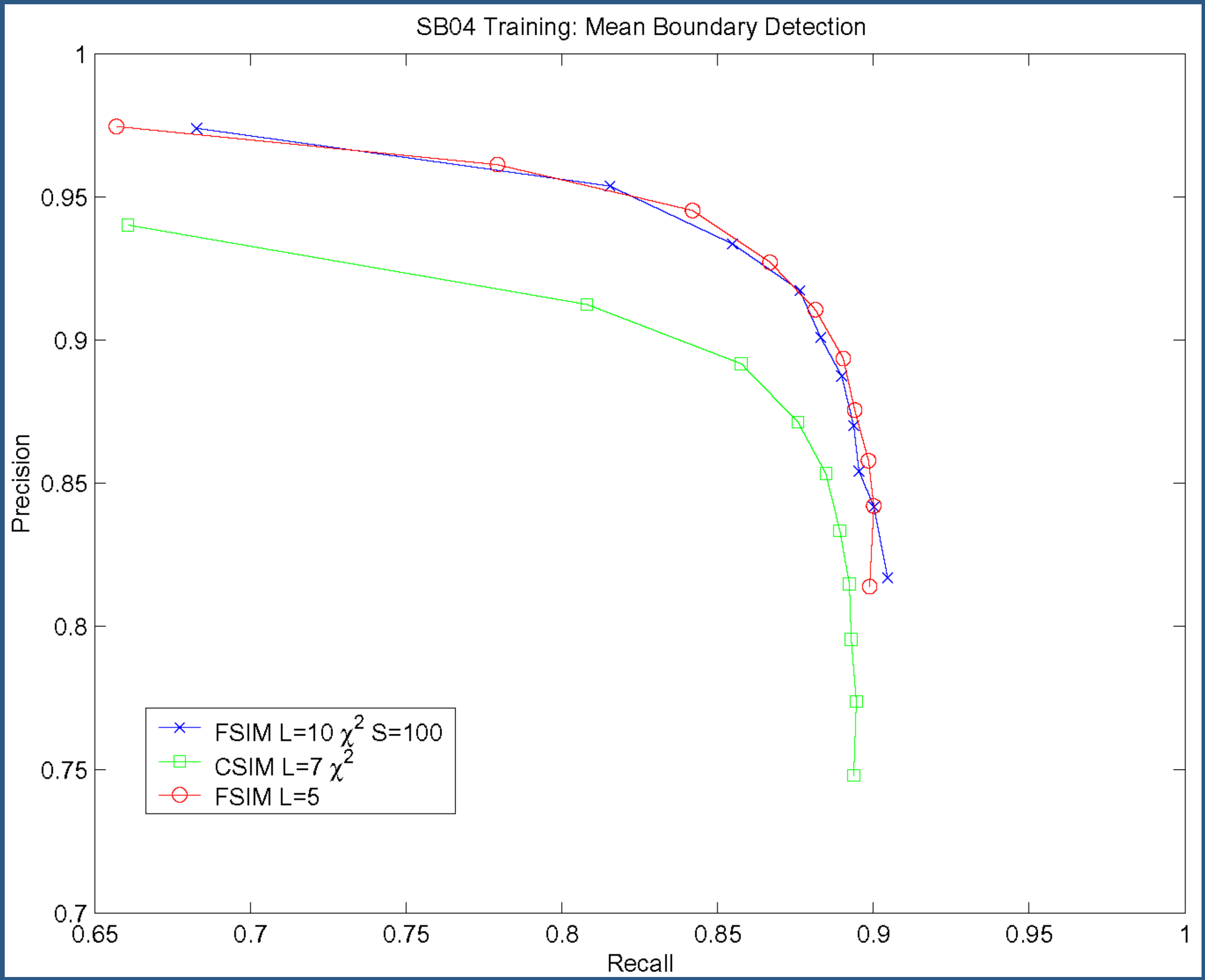
$$d' \geq 4(\epsilon^2/2 - \epsilon^3/3)^{-1} \ln N \qquad (1)$$

Then for any set V of N points in R^d, there is a map $f\colon R^d \to R^{d'}$ such that for all $u, v \in V$,

$$(1-\epsilon)||u-v||^2 \leq ||f(u) - f(v)||^2 \leq (1+\epsilon)||u-v||^2. \qquad (2)$$

  - easy to implement: O (d'dN)

varying *L* for fixed featured dimensionality



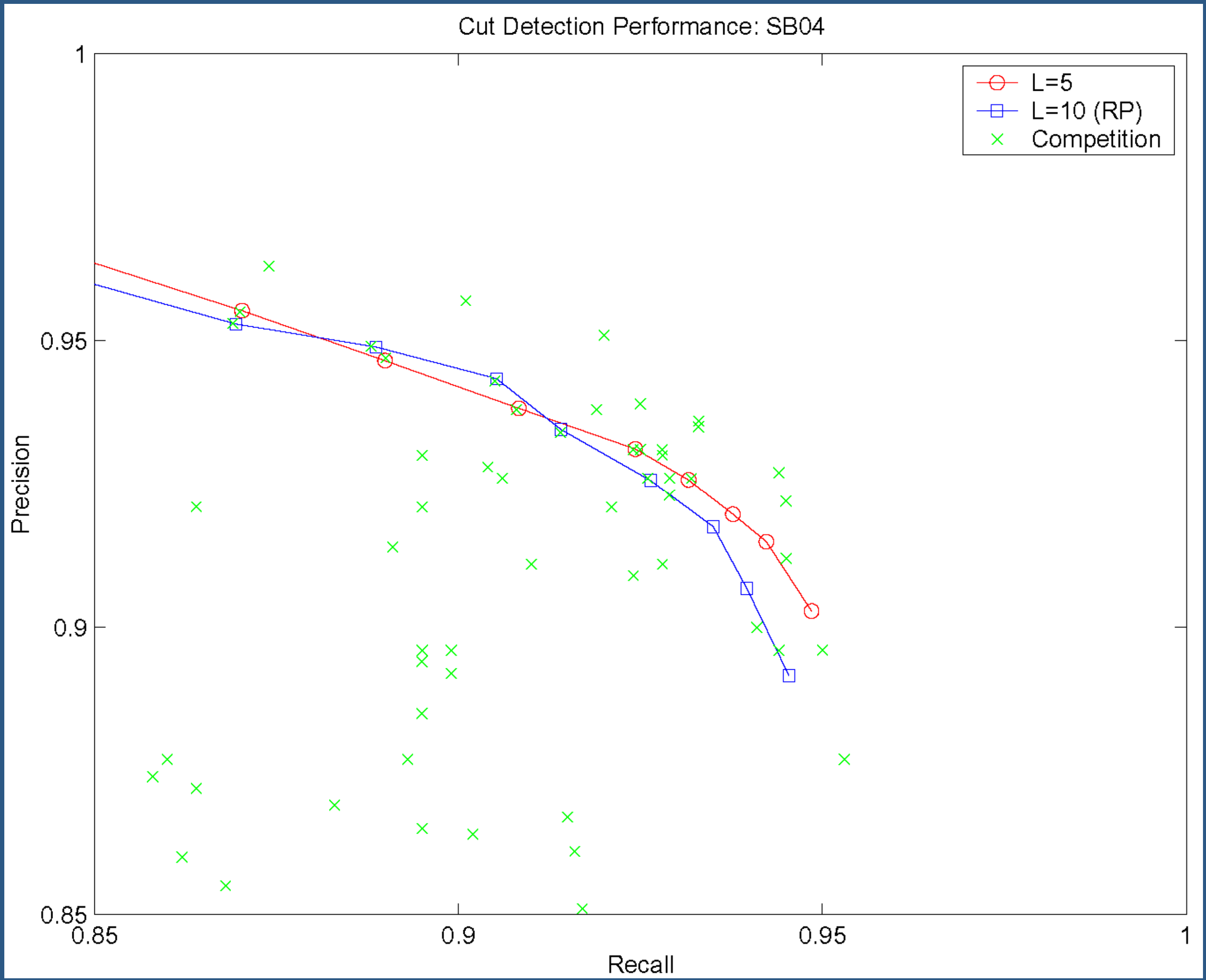
SB04 systems

- training data consists of 8 ABC, CNN videos from SB03 set
- 90% of non-boundary frames discarded
- k = 11
- sensitivity determined by $0 \leq \kappa \leq k$

- post-processing to avoid spurious boundaries in local temporal neighborhood

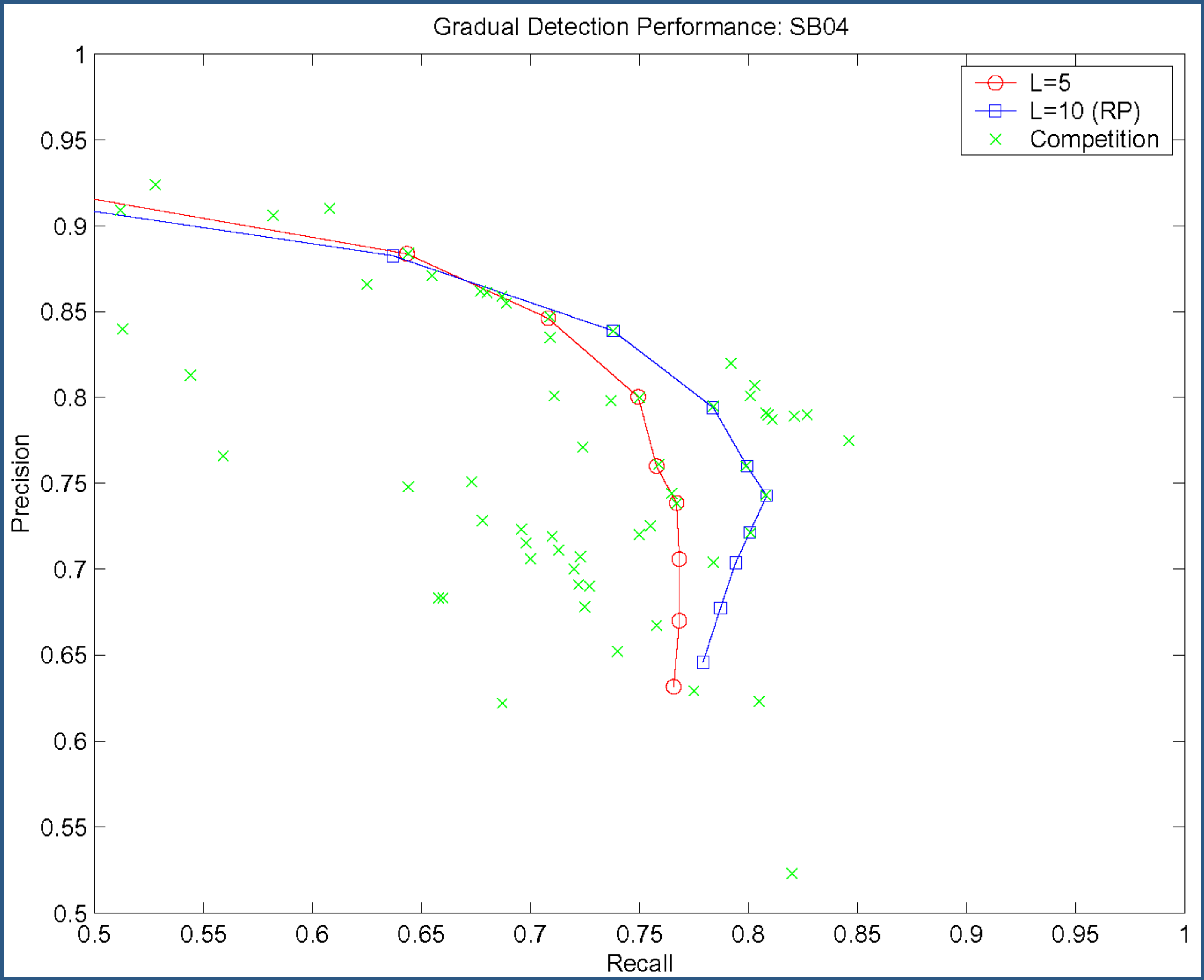
Cut Results

	R	P	F
Avg	0.831	0.762	0.776
Best	0.920	0.951	0.935
<FXPAL>	0.903	0.940	0.921



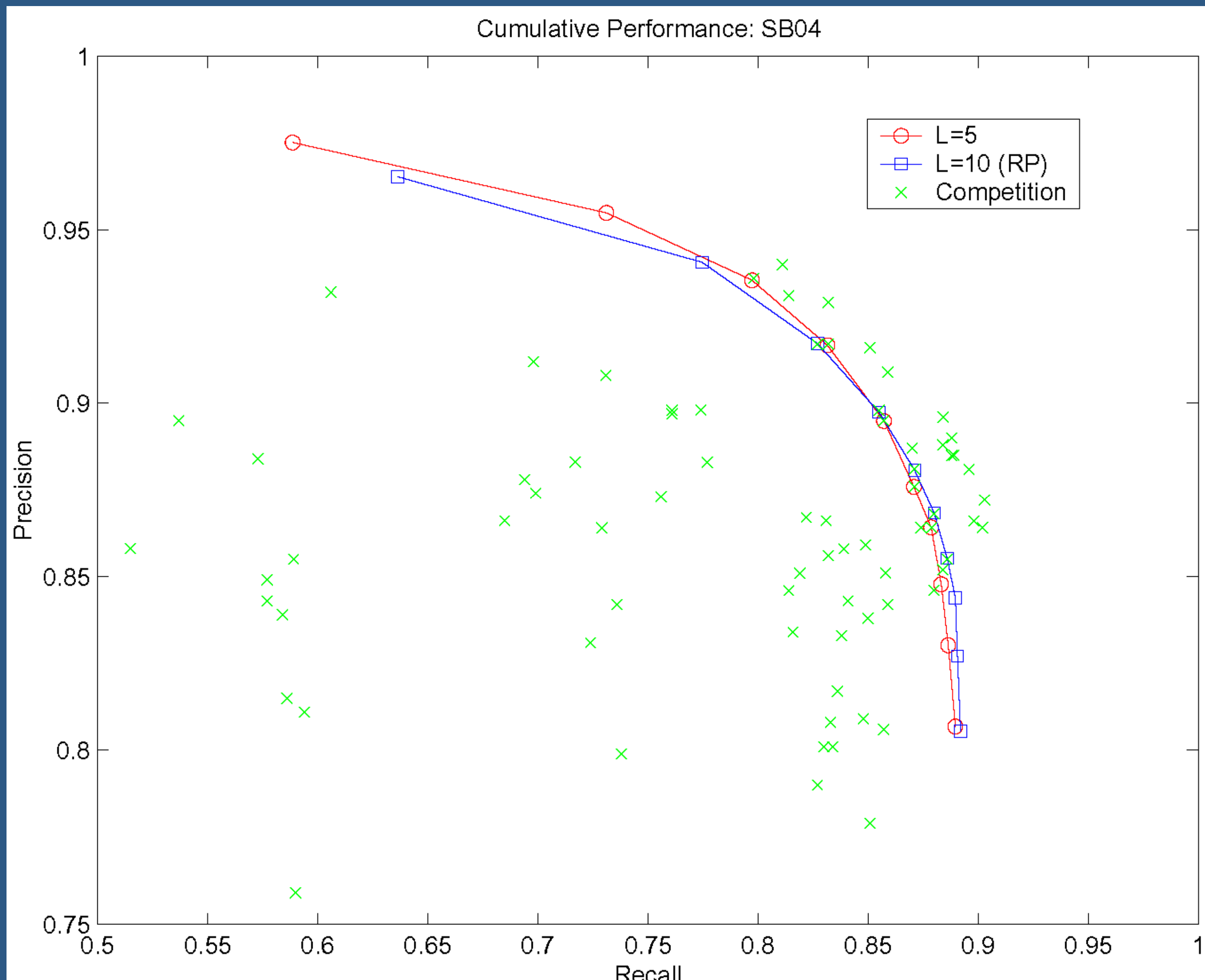
gradual results

	R	P	F
Avg	0.503	0.578	0.565
Best	0.846	0.775	0.8089
<FXPAL>	0.756	0.789	0.769



mean results

	R	P	F
Avg	0.7255	0.727	0.709
Best	0.884	0.896	0.890
<FXPAL>	0.856	0.891	0.872



time complexity

SysID	Decode/Extract	kNN	PostProcess	TOTAL	Ratio to Real Time
FS05_04	24882.350	20183.000	7.800	45073.150	2.087
FS05_05	24882.350	20183.000	7.789	45073.139	2.087
FS05_06	24882.350	20183.000	7.831	45073.181	2.087
FS05_07	24882.350	20183.000	7.831	45073.181	2.087
FS05_08	24882.350	20183.000	7.870	45073.220	2.087
FS10_04	24882.350	21825.000	7.811	46715.161	2.163
FS10_05	24882.350	21825.000	7.793	46715.143	2.163
FS10_06	24882.350	21825.000	7.809	46715.159	2.163
FS10_07	24882.350	21825.000	7.801	46715.151	2.163
FS10_08	24882.350	21825.000	7.830	46715.180	2.163

- 1 decode run includes histogram extraction (code never optimized) for all SysIDs
- 2 classification runs correspond to 10 SysIDs
- all times for all 12 videos

conclusions

- many segmentation approaches can be formulated within the framework of inter-frame similarity analysis and linear kernel correlation
- non-parametric supervised classification is effective for media segmentation
- very general framework

- thanks to Andrew Moore at CMU
- for more information: cooper@fxpal.com