Robust Video Duplicate Detection & Localization in Very Large Repositories

Zhu Li

Dept of Computing

The Hong Kong Polytechnic University

Outline

- **Introduction**
- **Research Motivations**
- **A brief overview of my current projects**
  - Intelligent Video Networking
  - Mobile Search Services
  - Video Signal Processing
- **In depth discussion**
  - Likelihood prunning based video duplicates detection,
    - » Duplicate likelihood modeling,
    - » Likelihood fast approximation via multi-indexed mahalonobis distance evaluation.
    - » Fast prunning algorithm
  - Backup topics (prob for individual meeting) :
    - » Very Low Bit Rate Video Adaptation for Multi-Access Networks
    - » Mobile Location/Product Search
    - » Source Channel Coding in Video Broadcasting, DE-STC/Foutaint/Layer
- **Summary & Questions**

About Me: http://www.comp.polyu.edu.hk/~cszli

- **Bio:**
  - **Zhu Li**, Asst Prof, ***HK Polytechnic Univ***, 2008.04~to date.
  - Senior, Senior Staff, and then Principal Staff Researcher, Multimedia Research Lab, ***Motorola Labs***, USA, 2000-08.
  - Software Engineer, CDMA Network Software Group***, Motorola CIG***, USA, 1998-2000.
  - PhD in Electrical & Computer Engineering, ***Northwestern University***, USA, 2004.
  - IEEE Senior Member
  - Vice Chair, IEEE Multimedia Communication Tech Committee, 2008~2010.
- **Research Interests:**
  - Video Coding and Adaptation, Optimization and Distributed Computing in Video Networking with applications in mobile TV, wireless video on-demand streaming, and P2P video networking.
  - Image/Video Analysis, Machine Learning and applications in Scalable large video repository search and mining problems.

2010

Devices

- **Explosive growth of devices:**
  - Billions of cell phones/PDAs
  - Billions of computers
  - Billions of TVs
  - Billions of Media Players

- **Different Multimedia Capabilities in:**
  - display,
  - capture,
  - storage,
  - computing,
  - communication

Networks

- **Better technology from equipment makers**
  - Better wireless spectrum efficiency, WiMAX/LTE
  - High speed DLS/Cable, 100x100
  - Fiber optical solutions, GPON
- **More capacity from service providers**
  - More bandwidth, better coverage,
  - Convergence of data, voice and media service from service providers
  - Vertical integration of application and services

Content

- **Explosive growth of digital media**
  - Web, Email, Audio, Video, Game
  - News, Music, Movie, Talk show, Game, 2nd Life.
- **Rapid changes in the way contents are produced and consumed**
  - Personal vs Commercial
  - Passive (TV) vs Interactive (Blog, Game)
  - Centralized vs P2P


People's Need

- **People's need:**
  - Good Access, be able to get what you want, a storage and communication problem
  - Mobility across devices and access sessions: anywhere, on any device, not tied to a single device/location, get what they want, with good media quality (coding) and availability (communication/networking).
  - Intelligence and Personalization: be able to find what they are interested in and locate what they want, browsing with (implicit and explicit) personal preference.
  - Self-expression, Interaction and Social Networking, P2P video,video blog, live events streaming, social group based video sharing. Immersive video interaction.

Technology

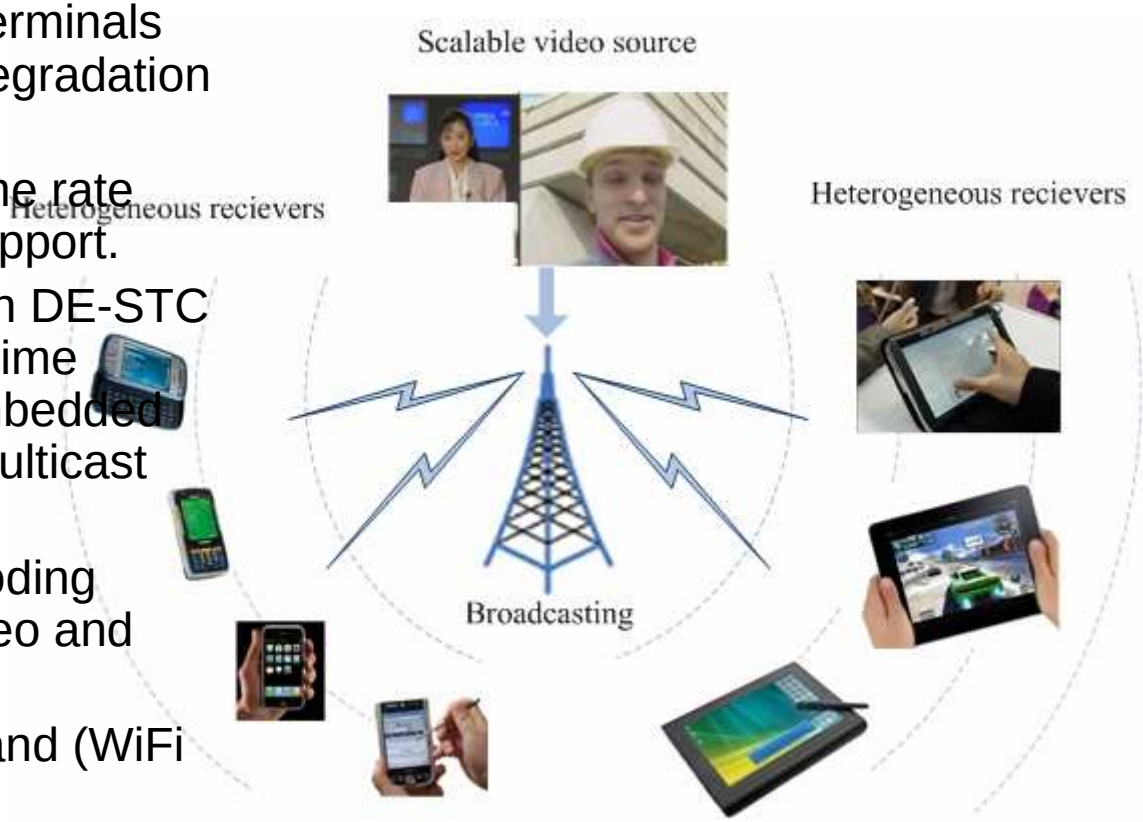
Technology Gap ?

- Distribution/Storage:
  - » With the popularity of smart phones and video applications, wireless networks are showing sign of strain.
  - » Internet re-engineering to support video dominating traffics
  - » Storage of multimedia content in cloud – explore various error-resilience and rate-distortion tradeoff characteristics to enable differential storage service.
- Search & Mining,
  - » Web scale multimedia analysis, indexing and retrieval,
  - » Integration with mobile applications.
- Interaction,
  - » New sensors: visual/audio/motion/supersonic wave sensors and processing,
  - » New algorithms: pattern recognition, visual tracking, immersive video.

Current Research Projects Highlights

Project Highlights – Video Communication

- *Intelligent Mobile TV*
  - Supporting highly elastic and robust QoS for a variety of mobile terminals with graceful visual quality degradation with channel conditions
  - Practical frame-size and frame rate adaptation for wide codec support.
  - PHY layer adaptation through DE-STC (diversity embedding space-time coding) to induce a set of embedded channels that best suit the multicast group channel distribution
  - APP layer source-channel coding optimization with layered video and digital fountain code
  - Targeting in-band, or dual-band (WiFi + 3G) wireless infrastructure.



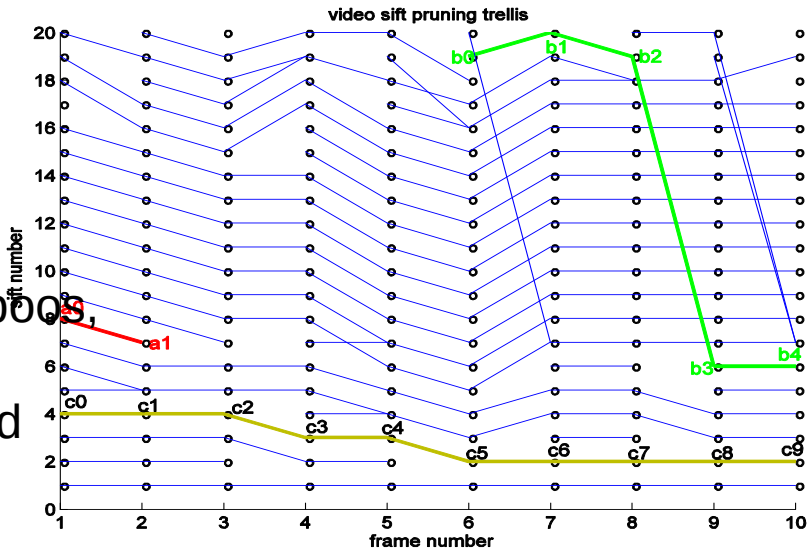
Project Highlights – Video Communication

- *Next Generation Content Networks*
  - Video accounts for > 70% of internet traffic now
  - Optimization and distributed computing solutions in IPTV video networking Primal-Dual decomposition, resource pricing schemes for multi-access and IPTV video networks, (RGC New Staff Grant, in collaboration with EDGE lab)
  - Video TCP: TCP re-engineering for content delivery networks. Reconsider the congestion measure and pricing as well as source adaptation schemes in TCP to better suit for content intensive traffics.
  - Caching and Network Coding schemes for video sharing in Mesh Networks (in collaboration with Prof Cao). Also integration of routing and flow control in (small scale) mesh networks for video multicasting.

Project Highlight - Mobile Search

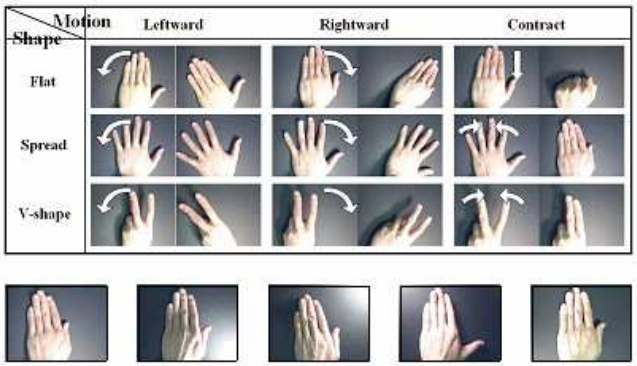
- ***Video Based Mobile Location Search*:**
  - Location search by mobile video capture and query
  - Video SIFT points indexing with appropriate scale and spatio-temporal quality metrics
  - Fast search with multi-indexing of SIFT point sets.
  - See my recent ACM MM paper for more detail.

- ***Image/Video Based Mobile Product Search* :**
  - Query-by-Capture, when shopping in the malls, just took a picture/video of the product and search the online stores to compare prices.
  - Crawled more than 1 million images from taobao.com
  - SIFT/local color feature based indexing
  - Works well for certain categories, like shampoos, shoes, …etc
  - Key technical challenge – offline learning and web based labeling.



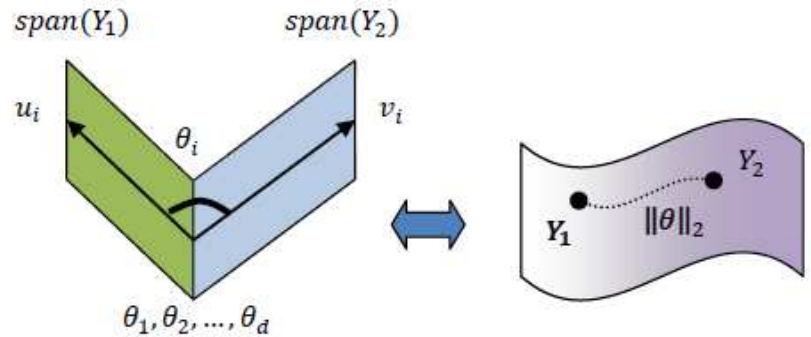
Project Highlights – Video Analytics

- ***Video Duplicate and Near Duplicate Search and Mining***:
  - Robust and scalable video fingerprint for duplicate and near duplicate search and mining applications (Funded by Microsoft Research grant, HK RGC GRF grant)
  - Will be discussed in more detail.
- ***Video Action and Event Recognition*:**
  - Video action and human behavior recognition based on spatio-temporal appearance volume manifold modeling, exploring tensorial, spline approaches, and novel learning solutions, e.g, aligned projection, HMM, and DBN. (Funded by HK RGC and PolyU internal grant)
  - Target Apps: surveillance, video biometrics, video search, social networks



Project Highlights – Video Analytics

- ***Subspace Indexing on Grassmann Manifold***:
  - For a large subject set pattern recognition problem, single subspace model's DoF is not enough for robust recognition
  - Instead, develop a rich set of transforms that better captures local data characteristics, and
  - Develop a hierarchical index for subspaces on the Grassmann manifold.
  - Target applications: large subject set face recognition, hierarchical transforms for image coding.



Robust Video Duplicates Detection & Localization in Large Video Repositories

Zhu Li

Dept of Computing
The Hong Kong Polytechnic University

- **Outline**
  - The duplicate and near duplicate localization problem
    - The problem
    - The applications
    - The challenges
  - Duplicate Likelihood Maximization Formulation
    - The duplicate likelihood Gaussian Process modeling
    - Duplicate Likelihood approximation via multi-indexed locality search
    - Sequence Likelihood pruning
  - Simulation Results & Discussions
    - Data Set and Simulation Setup
    - Accuracy, Complexity and Tradeoffs
  - Conclusion, Related & Future Work

The Application & Problem

- **Robust duplicate and near duplicate video detection and localization has many applications**
  - Copyright protection
  - Video repository mining – find out how video programs are related
  - Query by capture
- **The challenges:**
  - ***Robustness***:
    - Be able to detect edited,corrupted and reformatted video duplicates.
    - Be able to locate the duplicate clips in a very large repository
    - Scales with the length of the query clip
  - ***Complexity***:
    - The algorithm should scale well with the size of the repository,
    - Complexity – Robustness tradeoffs
    - Parallelization

The Duplicate Likelihood of a frame

- **Duplicate frames are not the exact copy of the original in the repository**
  - Otherwise, the problem is becoming trivial, just design a hash function would give us the exact match
- **Instead, duplicate frames can have various degrees of degradations and corruptions due to:**
  - Coding and Communication Losses, i.e, quantizaiton effect, packet losses
  - Image formation variations: lighting change, re-capture process induced affine transforms, ...etc
  - Editing effects: subtitles, graphics and text overlay, re-sizing, cropping...etc

The Duplicate Likelihood of a frame

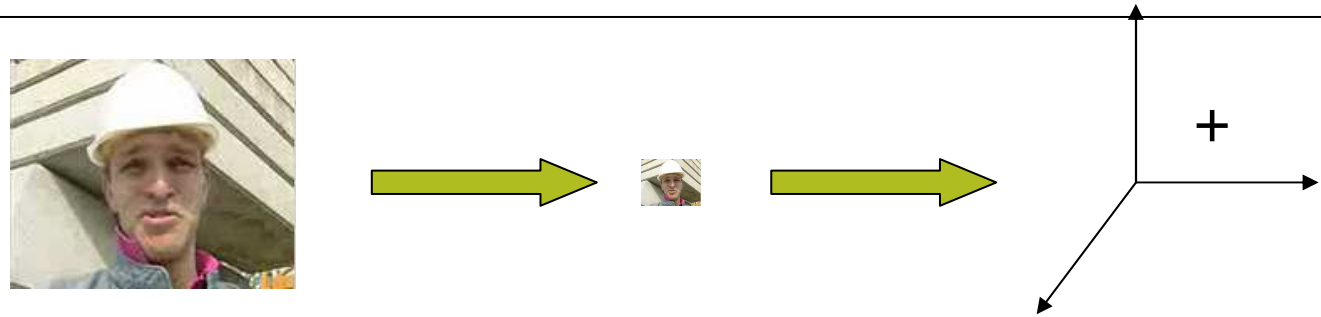
- **This can be captured by a probabilistic model, for a given frame x, the likelihood of y is a duplicate of x is given by a Gaussian distribution:**

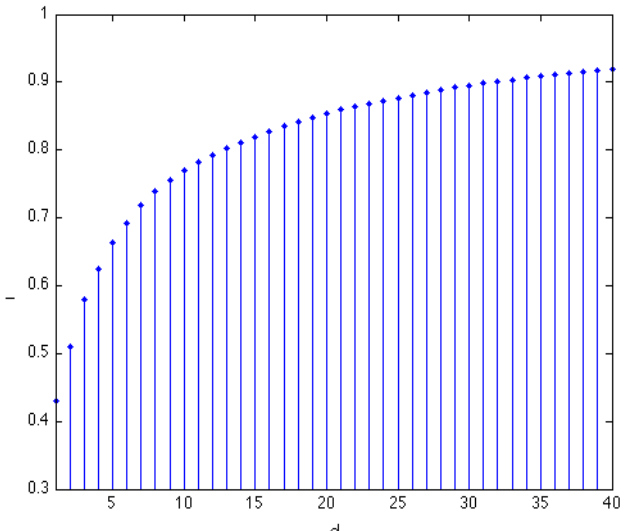
$$\mathcal{L}(y;x) \sim \mathcal{N}(m_x, \sigma_x)$$

$$\mathcal{L}(y;m_x,\sigma_x) = \frac{1}{2\pi^{d/2}|\sigma_x|^{1/2}} e^{-\frac{1}{2}(y-m_x)^T\sigma_x^{-1}(y-m_x)}$$

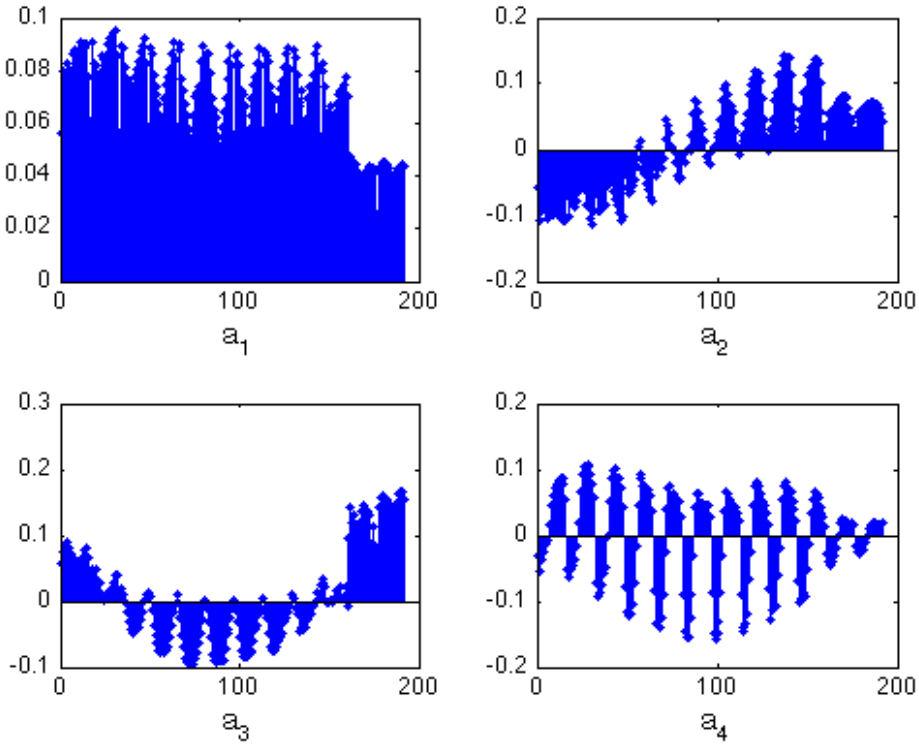
  - With d x1 mean vector and dxd variance matrix in feature space of choice
- **The solution is actually feature agnostic.**
  - The choice of feature space is up to the application, and the amount of loss, corruption and editing effects in the video frames affects the covariance matrix.
  - The choice of frame vs shot/GoP level features only affects the localization accuracy in this framework.

An Scaled Appearance Feature





Info loss vs dimensions

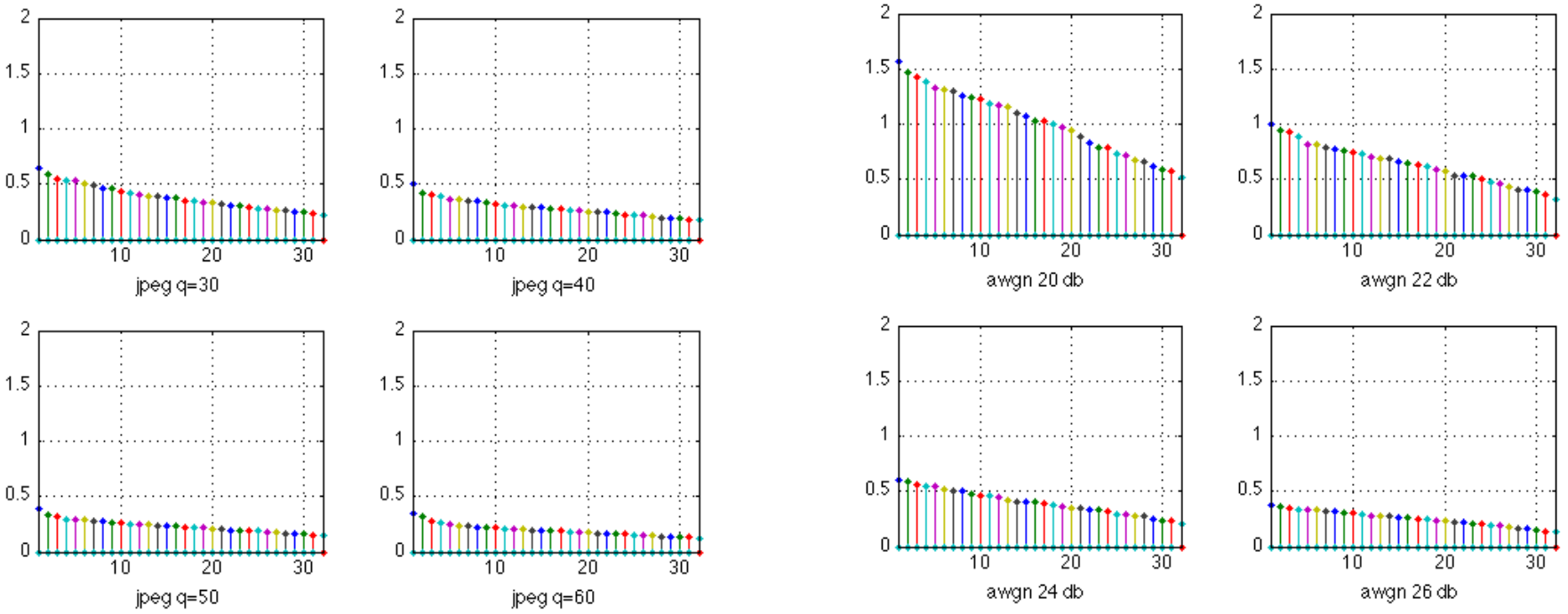


Basis functions

- **Icon frame and its PCA projectio**
  - Video Frame size: W=352, H=288
  - Icon size: w=16, h=12
  - A: d x 192,

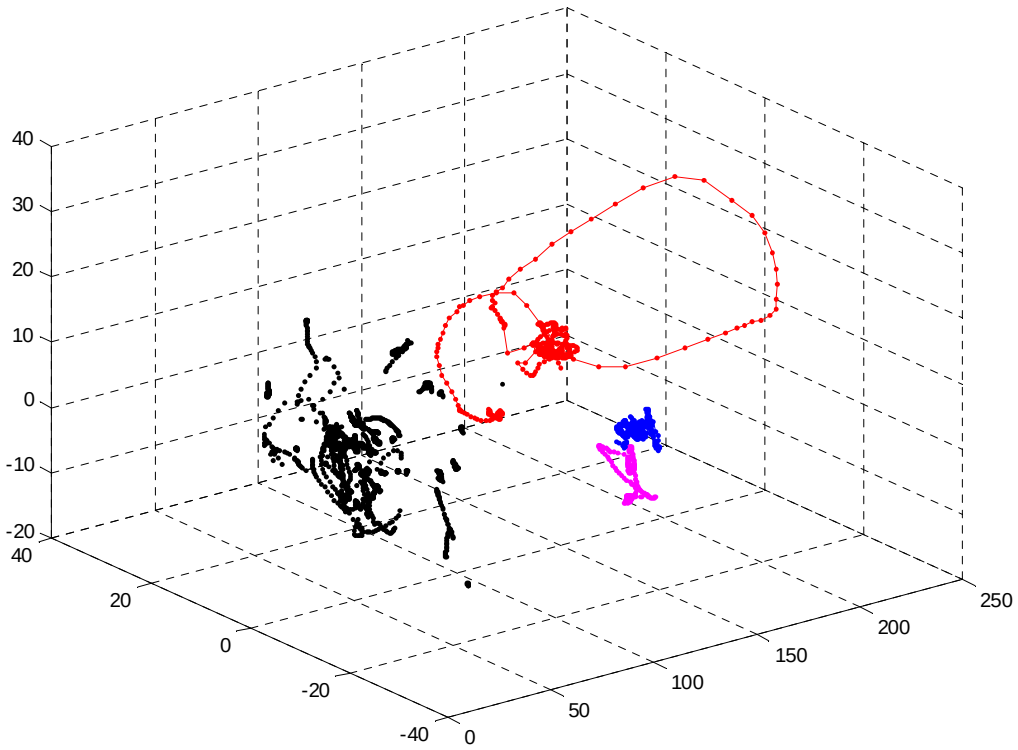
Likelihood Covariance Modeling

- **The mean of duplicate likelihood function is the original frame in the feature space**
- **The covariance is determined by the type of feature and severity of the degradation of the duplicate frame, examples of JPEG quantization and AWGN cases, 32x32**



Likelihood mean process : x(t)

Video sequence mean process $m_x(t)$ in PCA space with 1st,2nd and 3rd components



- "foreman" : 400 frames
- "stefan" : 300 frames
- "mother-daughter": 300 frames
- "mixed": 40 shots of 60 frames each from randomly selected sequences.

. "**foreman**", . "**stefan**", . "**mother-daughter**" , . "**mixed**"

Duplicate Likelihood Function

- Given a video sequence x(t),t=1..n, the likelihood of a sequence y(t) is a duplicate of x(t) is given by,

$$\mathcal{L}(y(t);x(t)) = \prod_{t=1}^{n} \mathcal{L}(y(t);m_x(t),\sigma_x(t))$$

  - Where $m_x(t)$ is the mean and $\sigma_x(t)$ is the covariance matrix of the likelihood function at time t.

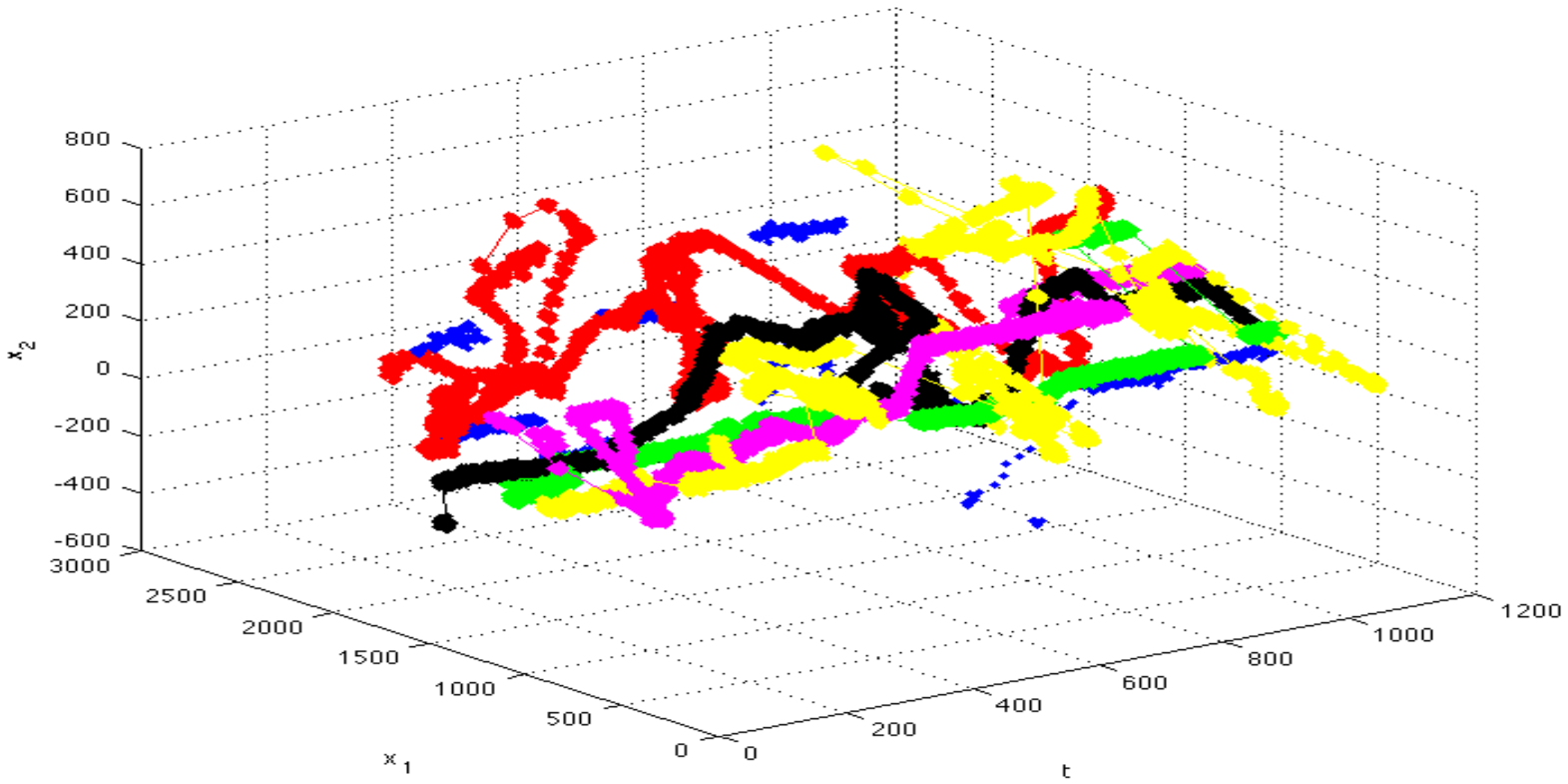
- Then the video duplicate detection problem is a maximum likelihood problem, given k=1..K video sequences { $x_k(t)$ } in the repository, the duplicate is found by maximum likelihood detection:

$$k^* = arg\max_k \prod_{t=1}^{n} \mathcal{L}(y(t);m_k(t),\sigma_k(t))$$

$$\begin{cases} m_k(t) = x_k(t) \\ \sigma_k^2(t) = \sigma_{quant}^2 + \sigma_{loss}^2 + \sigma_{gama}^2 + \sigma_{edit}^2 \end{cases}$$

Illustration of the Duplicate Likelihood Function for sequences

- **A set of 6 sequences and their likelihood function in $R^2(t)$: for a given query, need to find out which one gives the max likelihood.**
- **To un-tangle the mess.....**



Likelihood functions of 6 video sequences

The Likelihood Pruning Algorithm

- **Brutal force computing of likelihood function for all candidate sequences in the repository is computationally prohibitive**
  - Eg, if we have 1 million clips of 5 minute video, the total number of likelihood evaluation would be 10^6 x 5 x 60 x 30 = 9 billions of evaluations.
- **Instead, we can start with a set of sequences that sharing non-zero likelihood at certain timestamp, and pruning those sequences with zero likelihood at other timestamps.**
- **The key is to approximate the duplicate likelihood estimation, with a non zero-likelihood detection approximation via a multi-indexed locality search solution**
- **The sequence duplicate likelihood function is approximated by a time stamp subset pruning process, finding matching duplicate sequences as,**

$$S = \{k | \prod_{t \in ts} \mathcal{L}(y(t); x_k(t), \sigma_k(t) \geq 0\}$$

The Likelihood Pruning Algorithm

- This leads to a likelihood pruning algorithm:
  - Given query clip *y*(t), Compute a time stamp set, *ts*={t_1, t_2, …, t_m}
  - Init S_0= $\{k, |\mathcal{L}(y(t_1); x_k(t_1), \sigma(t_1)) \geq 0\}$
  - FOR i=1:m

$$S_{i+1} = S_i \bigcap \{k | \mathcal{L}(y(t_i); x_k(t_i), \sigma(t_i) \geq 0\}$$

    Break if S_i has 1 or 0 elements left.
  - END
- Criteria of *ts* selection:
  - The selection of time stamps should reflect maximum spatial-temporal diversity, in query clip y(t), i,e, find m points on y(t) that has the maximum separation possible among them, otherwise, becomes an image search problem

Timestamp Selection Heuristic

- **We want the selected matching point has the best spatial-temporal diversity, ie., don't want video duplicate search degrade into image duplicate search.**
- **An effective heuristic solution : Curve Length Guided Selection**
  - Compute cumulative curve length of y(t) as by fitting a spline model y(s) on y(t):

$$L_y(t) = \int_{s=0}^{t} y(s)ds$$

  - If to have m time stamps, just find step size as

$$ts(k) : L_y(ts(k)) = k\Delta, k = 1..m$$

$$\Delta = \frac{L_y(n)}{m}$$

Likelihood approximation via Multi-Indexing Locality

- We need to find out frames that have non-zero likelihood of a query frame y(t)

$$S(y) = \{x | \mathcal{L}(y; m_x, \sigma_x) \geq 0\}$$

This can be approximated by an epsilon-NN on m_x, i.e,

$$S(y) = \{x | d_{\sigma_x}(x, y) \leq d_0\}$$

- The choice of d_0 reflects the zero likelihood prob cut-off Mahalanobis distance.

$$d_{\sigma_x}(x, y) = (x - y)^T \sigma_x^{-1} (x - y)$$

Likelihood approximation via Multi-Indexing Locality

- This can be approximated by multi-indexing structure by searching for leaf node containing frame y:

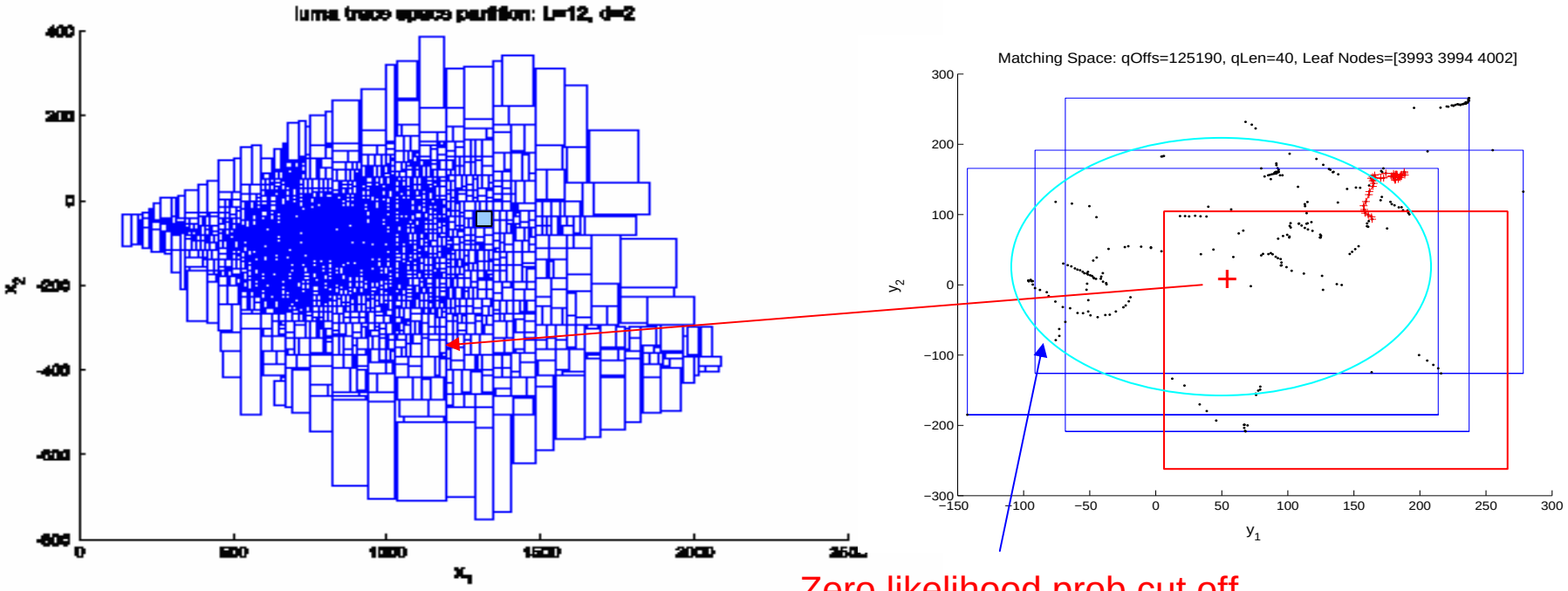
$$S(y) = \bigcup_j F_j(y)$$

where F_j(y) returns the indexing leaf node frame set from index structure j. By choosing an appropriate indexing height we can control the leaf node volume and therefore d_0.

- However, single indexing structure is not a good approximation: if query frame y is not at the centroid of the leaf node, then many non-zero likelihood frames will be missed.

Multi-Index Likelihood Approximation

- **Graphical illustration**
  - A query frame y, in +, is at northwest corner of its leaf node, $F_1(y)$ in index scheme 1. non-zero likelihood frames out side the red rectangles will be missed
  - But other 3 indexing scheme leaf nodes in black, can cover this loss
  - Careful selection of indexing scheme and leaf node volume can have a good approximation of the likelihood function



Likelihood Pruning

- Taking advantage of temporal dimension constraint, the sequence duplicate likelihood $L_k(\mathbf{y})$ becomes zero, if any frame's duplicate likelihood is zero:

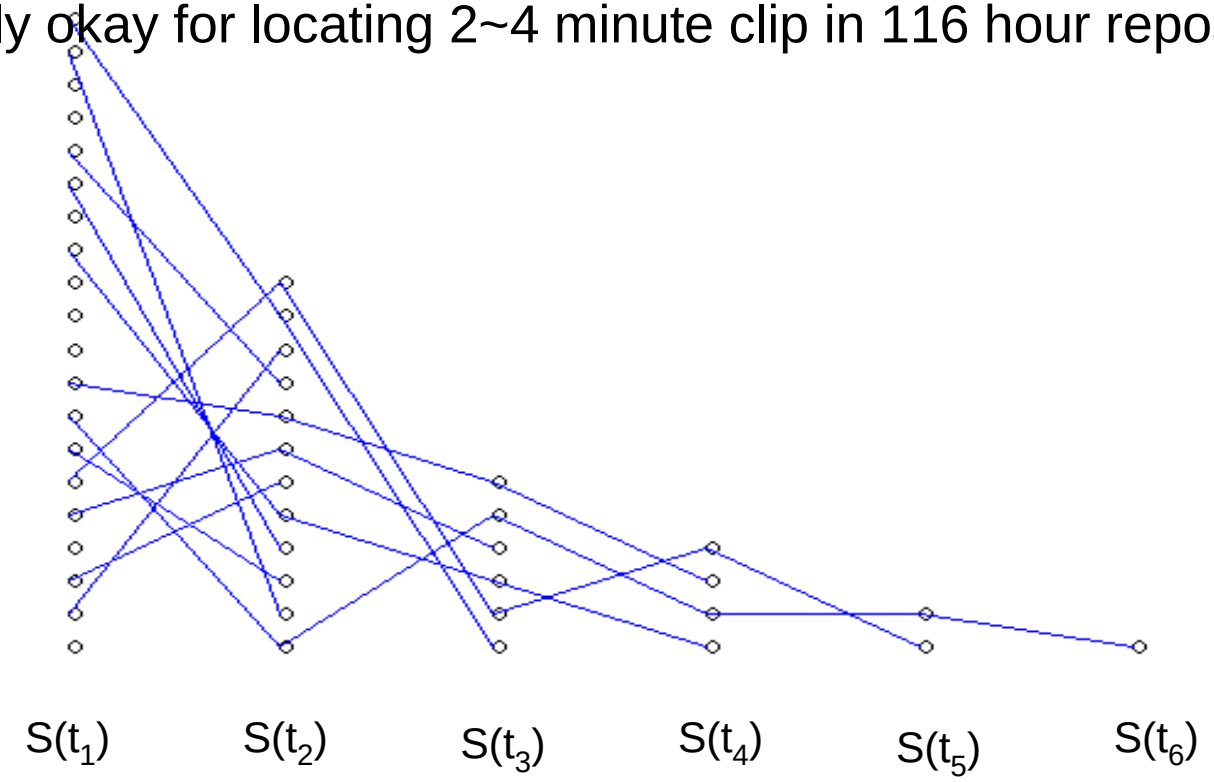
$$\mathcal{L}_k(\mathbf{y}) = \prod_{t=1}^{n} \mathcal{L}(y(t); m_k(t), \sigma_k(t))$$

$$\mathcal{L}_k(\mathbf{y}) = 0, \text{if} \exists t s.t. \mathcal{L}(y(t); m_k(t), \sigma_k(t)) = 0$$

- This means that at certain timestamp, t_1, if the duplicate likelihood of $y(t_1)$ is zero w.r.t sequence *k* in the repository, we can remove it from the candidate set.

Likelihood Trellis Pruning

- **Illustration of the process:**
  - Start with initial non-zero likelihood set
  - Prune those having zero likelihood in the next stage of likelihood evaluation
  - Stop 'till only one, or no trellis left in the node (in simulation, 3~5 stages typically okay for locating 2~4 minute clip in 116 hour repository)

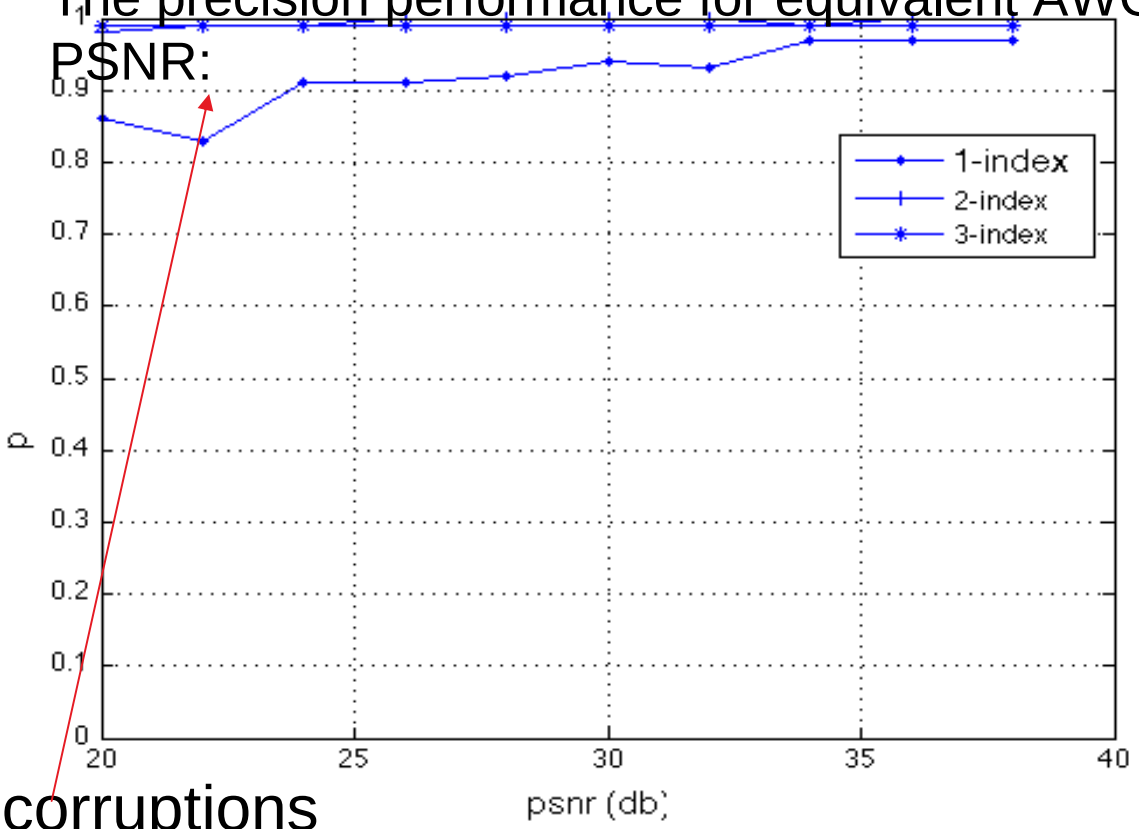


Simulation

- **Data Set**
  - Indexed Set:
    - 12 x 2^{20} frames, or approx 116 hours of video from various sources
    - Each frame is scaled to 16x12 icon size, and projected to a 32-dimensional space
    - 4 kd-tree index structure is built for 1~8, 9~16, 17~24, 25~32 dimensions, to provide diversity in approximate likelihood function
    - Alternative approach: hierarchical quadtree structure that can have better likelihood approximation.
  - Negative Probe Set:
    - 5 hours of video from TRECVID shot detection set
    - Scale to 16x12 icon and project to 32 dimensional space
  - Distortions/Image Formation Variations in query
    - Quantization induced distortions: JPEG quality 30, 40, ..., 90.
    - Block losses: random block loss of rate 2%~10%
    - Blurs, Gama Corrections
    - AWGN: equivalent to 20dB~40dB PSNR range

Performance - Accuracy

- Randomly generated 4000 negative probes and 8000 positive probes with various degree of corruptions
- Probe clip length = 1, 2 and 4 minuets
- Recall = 100%.
- The precision performance for equivalent AWGN range of 20~40 dB in PSNR:



Diversity in likelihood Modeling improves accuracy

Bad corruptions

Complexity Analysis

- **Duplicate Likelihood Estimation**
  - Basically, O(L), for repository with a L level kd-tree.
    - Notice that if we allow sacrifice in duplicate localization accuracy, a GoP level feature indexing can reduce the problem set size by 15 times.
  - If we involve multiple indexing structure, need to merge multiple localities to come up with non-zero likelihood repository frame estimate for the given probe.
  - The robustness of the likelihood estimation is tied to the index tree height, covariances of the distortions in the probe
- **Likelihood Pruning**
  - The pruning stops if there is only one trellis left, or no trellis.
  - Starting with an initial set, the size of candidate trellis decreases
  - Complexity is content dependent.
- **Parallelization**
  - The index search can be parallelized, as each index structure is independent.
  - Pruning

Performance - Complexity

- **Computing resource**
  - Pure Matlab implementation, not optimized thoroughly.
  - 3.0GHz Linux running Lenovo ThinkCenter, 4GB memory
- **Index complexity:**
  - To build $12x2^{20}$ 8-dimension kd-tree of 2^{14} leaf nodes, it takes approximately 63~65 seconds on a Linux PC with 3.0GHz processor.
  - Leaf node has 768 frames.
  - There is a tradeoff between the size of leaf nodes and the robustness of the likelihood estimation

Performance - Complexity

- **Likelihood evaluation complexity, 5 time stamps:**
  - For a given duplicate frame, to find all its non-zero likelihood repository frames takes average 0.13 ms for one index structure
  - For multi-index structure, find and merge non-zero likelihood repository frames takes approx 1.01 ms for 2 index case, and 2.20 ms for 3 index case.
  - Not much precision performance gain for 4 index case, so not reported.
- **Likelihood pruning complexity**
  - For 5 timestamp case, 0.67, 0.94, 0.94 ms for 1, 2, and 3 index structures
  - Not a significant source of complexity, can afford more elaborate pruning sequences
- **Total time complexity:**
  - Within 3 ms for 116 hour repository
  - Goal is to handle 10,000 hour within 10ms, will prob need more elaborate multiple pass likelihood pruning process

Conclusion & Future Work

- **Conclusion**
  - A feature agnostic and flexible likelihood modeling and pruning scheme for video (and also audio) duplicate detection and localization
  - Very high accuracy in precision-recall performance, roughly 98% precision on 100% recall, to localize 2~4 minute probes within 3 seconds in a repository of 116 hours
  - Good response time, takes only 3 ms for 116 hour repository and on track to achieve 10 ms response time for 10,000 hour repository (need some help in getting the data set).
  - Flexible Computational complexity – Robustness – Localization Accuracy scalability in performance.

Conclusion & Future Work

- **Future Work**
  - More sophisticated likelihood pruning schemes
  - Subspace learning in minimizing distortion induced likelihood covariance
  - A digital fountain decoding like solution, offers fine granular tradeoffs between complexity and accuracy.
  - Parallelization in Cloud.

Q &A

Thanks….