数字视频编解码技术国家工程实验室
National Engineering Laboratory for Video Technology

PKU-IDM @ TRECVID 2011 CCD:
Video Copy Detection using a Cascade of Multimodal Features & Temporal Pyramid Matching

Yonghong Tian

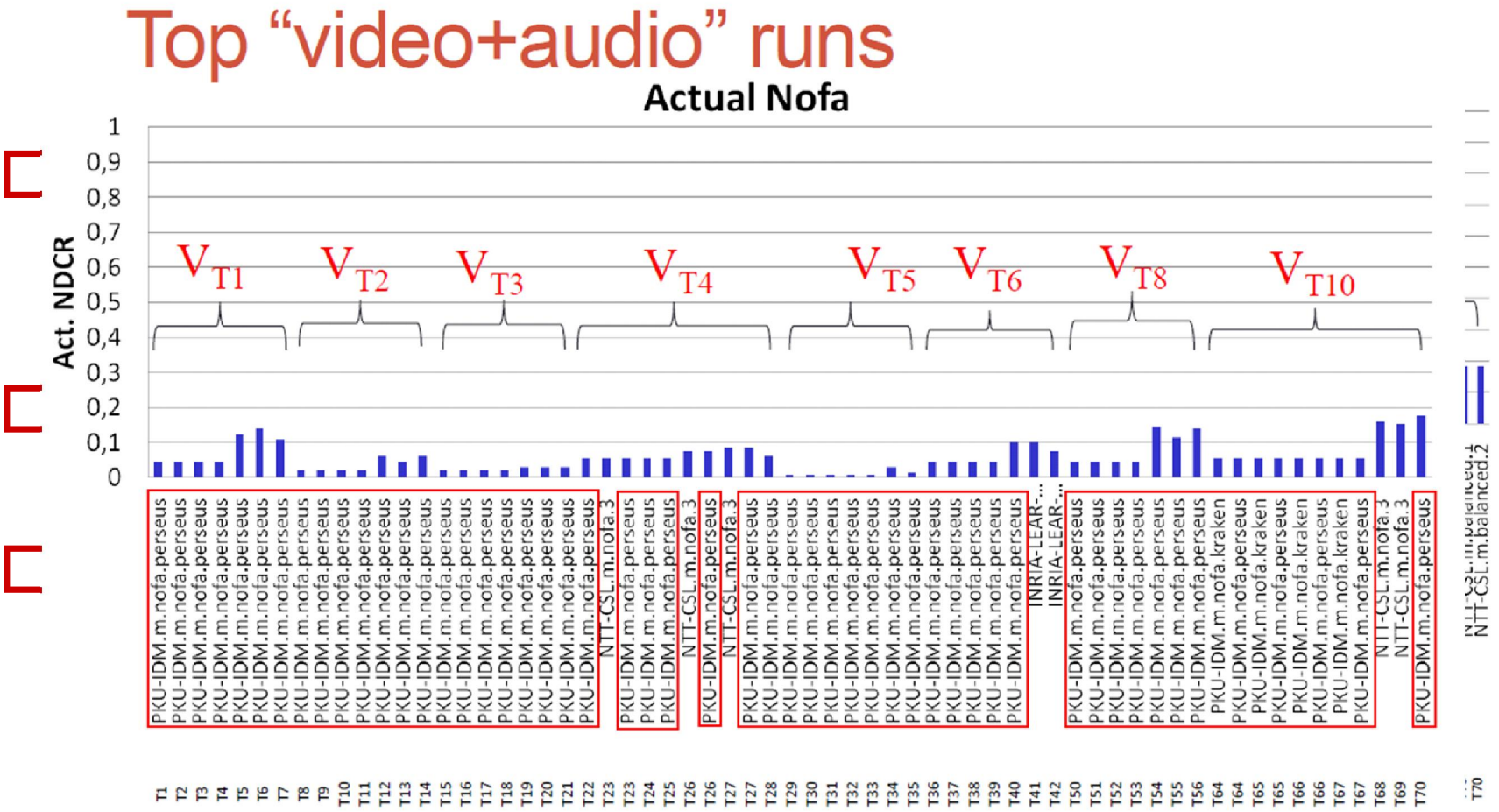
National Engineering Laboratory for Video Technology
School of EE & CS, Peking University

Outline

- Experience from CCD10
- Our Solution @ CCD11
  - Preprocessing
  - Complementary Multimodal Features & Indexes
  - Temporal Pyramid Matching
  - Cascade Architecture
- Evaluation Results
- Demo
- Summary

Experience from CCD10

- Our results @ CCD10

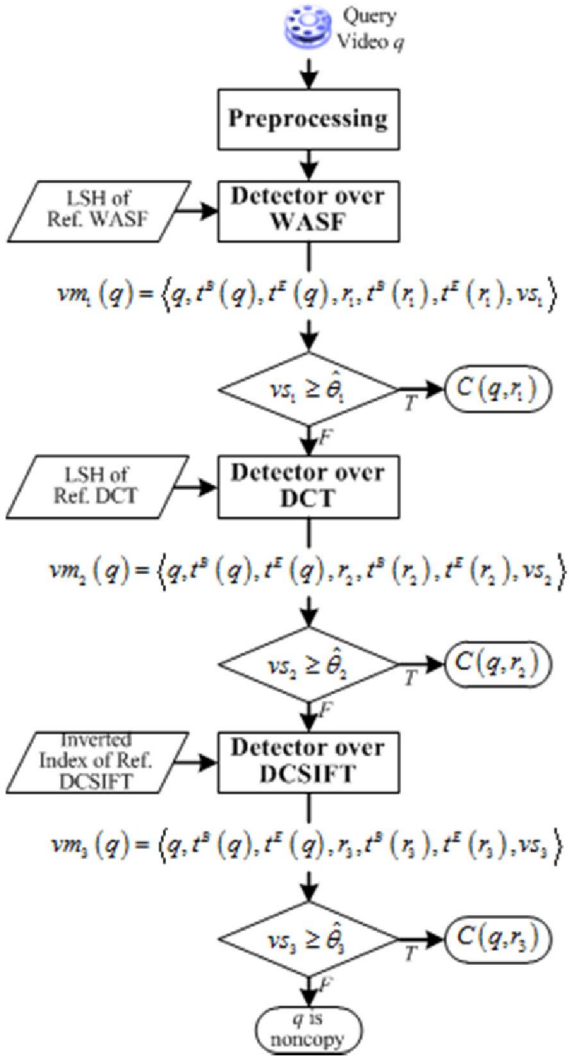


Experience from CCD10

- Strong points
  - Excellent detection effectiveness
    - Multimodal features
    - Temporal Pyramid Matching (TPM)
    - Preprocessing for PiP and Flip transformations
- Weak points
  - Bad efficiency
    - Redundancy of using SIFT & SURF simultaneously
    - Late fusion of results from all the basic detectors
    - Lack in parallel programming
  - Median localization accuracy
    - Overcautious strategy for copy extent computation in fusion module

Our Solution to CCD11

- Solution
  - Preprocessing
  - Complementary Multimodal Features & Indexes
    - DCSIFT BoW + Inverted Index
    - DCT + LSH
    - WASF + LSH
  - Temporal Pyramid Matching
  - Cascade Architecture
- Improvements from CCD10
  - DCSIFT instead of SIFT & SURF
  - Cascade architecture instead of Late Fusion & Verification



(1) Preprocessing

- Audio
  - Audio frame=90ms, overlap=60ms
  - Audio clip=6s (198 audio frames), overlap=5.4s
- Video
  - Uniformly sampled key frames (3 kf/sec)
  - Picture-In-Picture
    - Detect & localize PiP through Hough transform
    - Process foreground & original frames respectively
  - Flipping
    - Asserted non-copies will be flipped and matched again

(2) Complementary Multimodal Features

- What's *"complementary"*?
  - Basic assumption: none of any single feature can work well for all transformations.
  - Some features may be robust against certain types of transformations but vulnerable to other types of transformations, and vice versa.
- 1st Goal: ***Trade-off between effectiveness and efficiency***
  - DCSIFT: lowest NDCR, longest MeanProcTime
  - DCT / WASF: higher NDCR, much shorter MeanProcTime

Detector	Avg. NDCR	Avg. MeanF1	Avg. MeanProcTime
DCSIFT	**0.117**	**0.955**	249.636
SIFT	0.210	0.953	138.550
DCT	0.344	0.953	**6.381**
WASF	0.194	0.949	**5.486**

All experiments are carried on an Windows Server 2008 with 32 Core 2.00 GHz CPUs and 32 GB RAM.

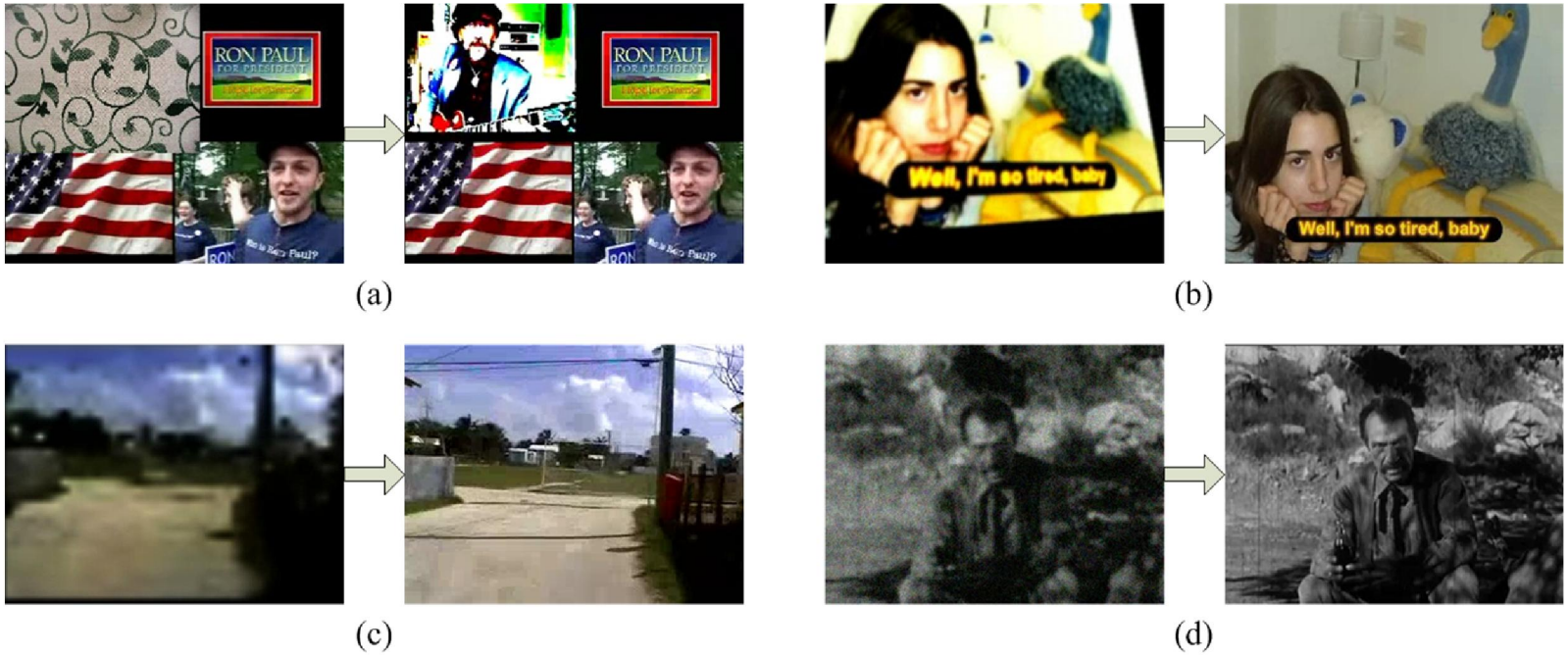
Complementary Multimodal Features

- 2nd Goal: ***Robust to different transformations***
  - DCSIFT / DCT vs. WASF
    - DCSIFT / DCT: visual transformations
    - WASF: audio transformations
  - DCSIFT vs. DCT:
    - DCT is more robust to severe blur and noise;
    - DCSIFT is more robust to other transformations

Detector	V1	V2	V3	V4	V5	V6	V8	V10	AVG
DCSIFT	**0.149**	**0.075**	**0.015**	0.104	**0.03**	0.261	**0.097**	**0.201**	**0.117**
SIFT	0.336	0.201	0.022	0.134	0.06	0.358	0.261	0.306	**0.210**
DCT	0.97	0.373	0.142	**0.097**	0.075	**0.224**	0.522	0.351	**0.344**

Complementary Multimodal Features

- Complementarity between DCSIFT and DCT
  - Only DCSIFT works
    - (a) V3-Pattern Insertion, (b) V1-Camcording
  - Only DCT works
    - (c) V6-Decrease in Quality (Severe blur), (d) V6 (Severe noise)



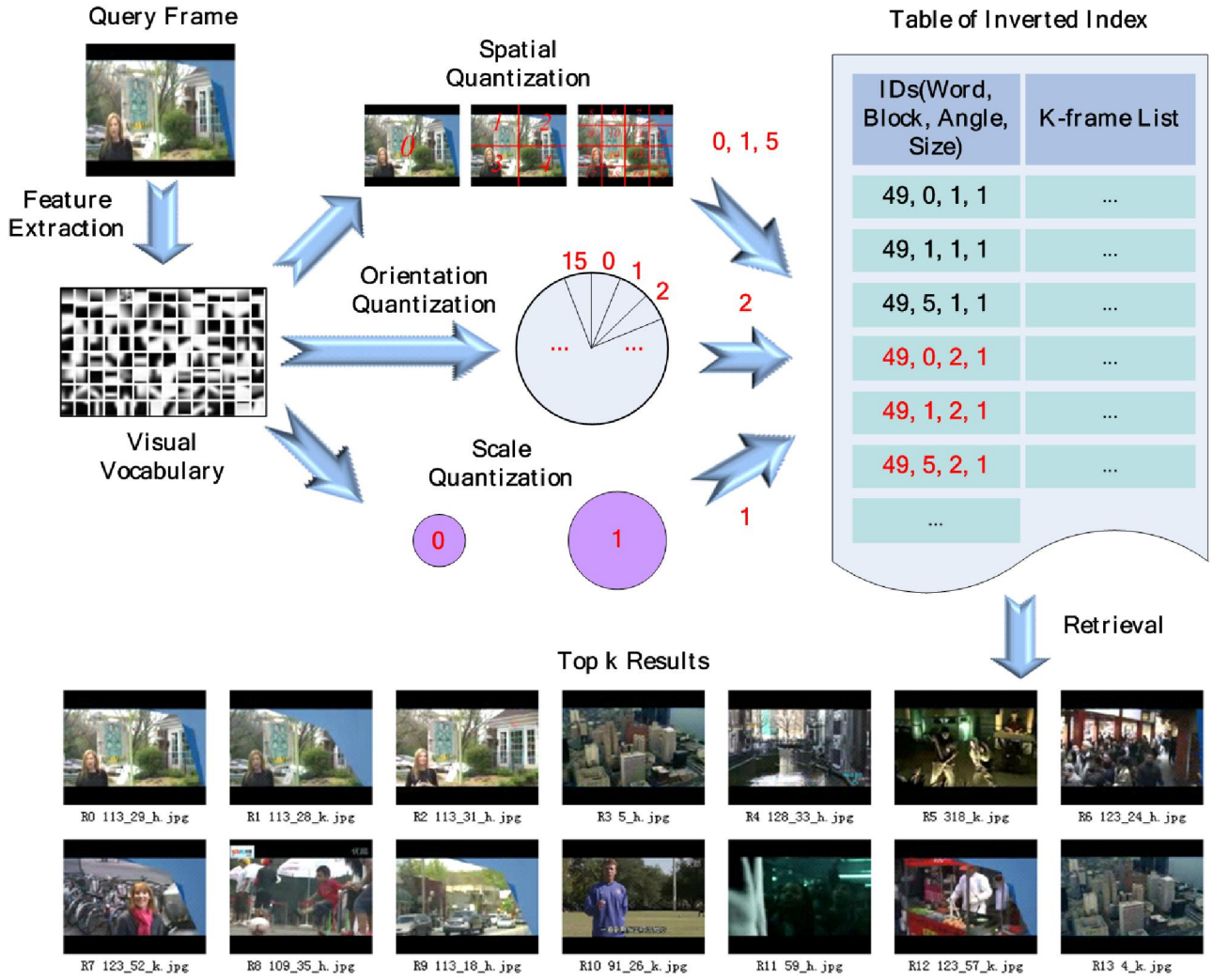
(a) DCSIFT BoW + Inverted Index

- Resist **content-altering** visual transformations
  - V1-Camcording, V2-PiP, V3-Pattern Insertion, V8-Postproduction
- Dense Color SIFT
  - Dense: multi-scale dense sampling instead of interest point detection
  - Color: sub-descriptors are computed from each LAB component and then concatenated to form the final descriptor
- BoW + Inverted Index
  - Use of position, scale and orientation
  - Enhance discriminability

Bosch, A., Zisserman, A., and Muoz, X. 2008. Scene classification using a hybrid generative/discriminative approach. IEEE Trans. Pattern Anal. and Mach. Intell. 30, 4, 712–727.

DCSIFT BoW + Inverted Index

- Key frame retrieval in DCSIFT detector

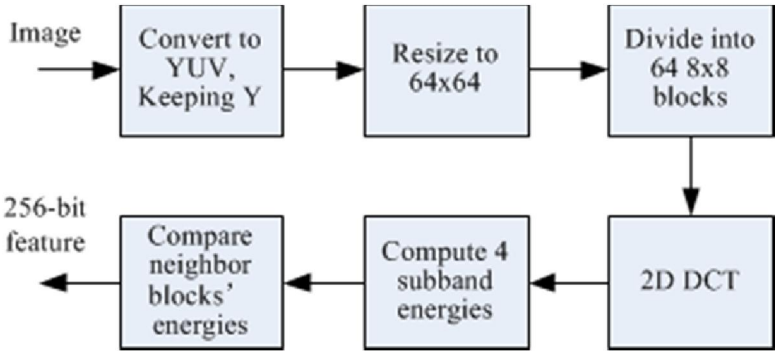


(b) DCT + LSH

- Resist **content-preserving** visual transformations
  - V4-Reencoding, V5-Change of Gamma, V6-Decrease in Quality
- DCT feature: DCT coefficient → subband energy

$$d_{i,j} = \begin{cases} 1, if\ e_{i,j} \geq e_{i,(j+1)\%64} \\ 0, otherwise \end{cases} \quad 0 \leq i \leq 3, 0 \leq j \leq 63$$

$$D_{256} = \langle d_{0,0}, \cdots, d_{0,63}, \cdots, d_{3,0}, \cdots, d_{3,63} \rangle$$

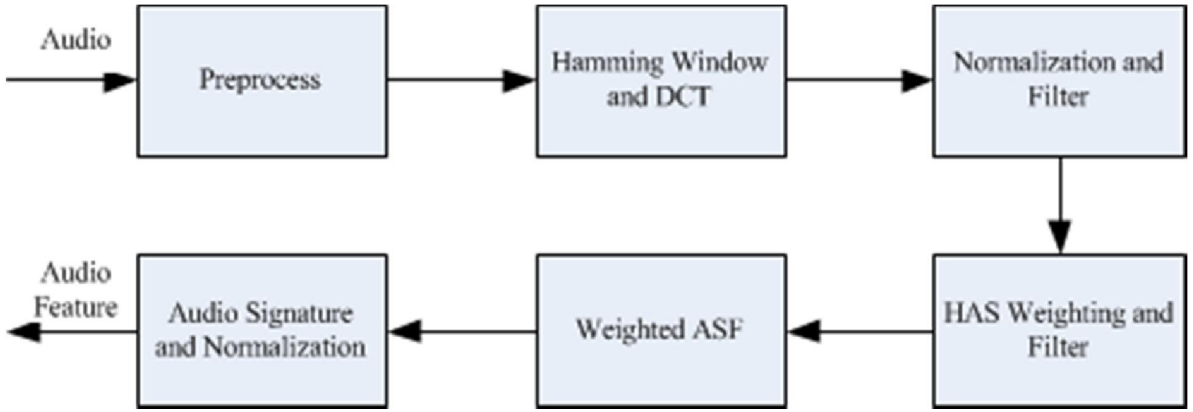


- Distance metric
  - Hamming distance
- Index
  - Locality Sensitive Hashing (LSH)

Subband 0	0	1	5	6	14	15	27	28
Subband 1	2	4	7	13	16	26	29	42
Subband 2	3	8	12	17	25	30	41	43
Subband 3	9	11	18	24	31	40	44	53
	10	19	23	32	39	45	52	54
	20	22	33	38	46	51	55	60
	21	34	37	47	50	56	59	61
	35	36	48	49	57	58	62	63

(c) WASF + LSH

- Resist audio transformations
  - A2-mp3 compression, multiband companding …
- WASF
  - To extend the MPEG-7 descriptor - Audio Spectrum Flatness (ASF) by introducing Human Auditory System (HAS) functions to weight audio data



- Distance metric: Hamming distance
- Index: LSH

(3) Temporal Pyramid Matching

- Temporal Matching
  - Integrate results of key frame (audio clip) retrieval into the result of video copy detection

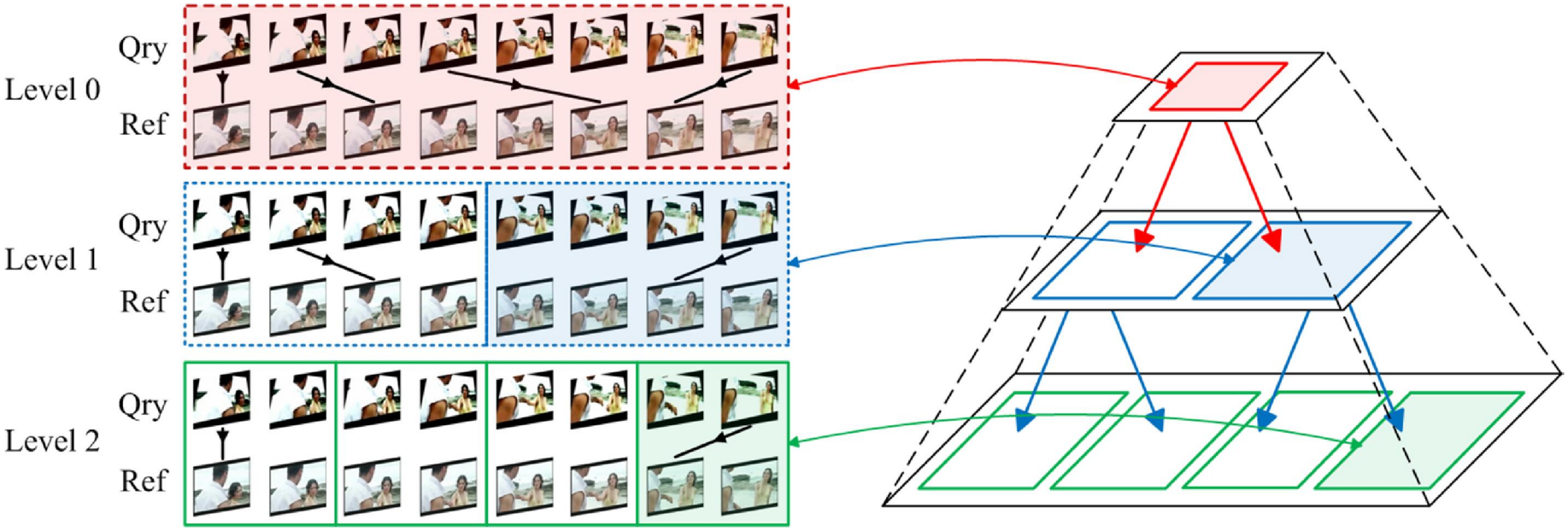
$$FM = \left\{ fm \mid fm = \left\langle q, t(q), r, t(r), fs \right\rangle \right\}$$

$$vm(q) = \left\langle q, t^B(q), t^E(q), r, t^B(r), t^E(r), vs \right\rangle$$

- Dilemma!
  - Matched frames between *q* and *r* should be aligned so as to eliminate mismatches
  - In practice, strictly aligned frame matches are so few, thus the above restriction might lead to more FNs

Temporal Pyramid Matching

- Key idea
  - Adapt "Pyramid Match Kernel" to 1-D temporal space
  - Partition a video into increasingly finer segments and calculate video similarities at multiple granularities



$$s_v = \kappa^L = 2^{-L} s_v^0 + \sum_{\ell=1}^{L} 2^{\ell-L-1} s_v^\ell .$$

Temporal Pyramid Matching

- Performance of DCSIFT detector with “TPM” vs. “Single Level Temporal Matching” on CCD09 and CCD10
  - TPM with a structure of four levels achieves the best matching result

ℓ	TRECVID 10		TRECVID 09	
	SINGLE LEVEL	TPM	SINGLE LEVEL	TPM
0 (1 ts)	0.273		0.219	
1 (2 ts)	0.247	0.223	0.192	0.179
2 (4 ts)	0.226	0.195	0.177	0.132
3 (8 ts)	0.202	**0.174**	0.173	**0.107**
4 (16 ts)	0.214	0.181	0.185	0.110

Temporal Pyramid Matching

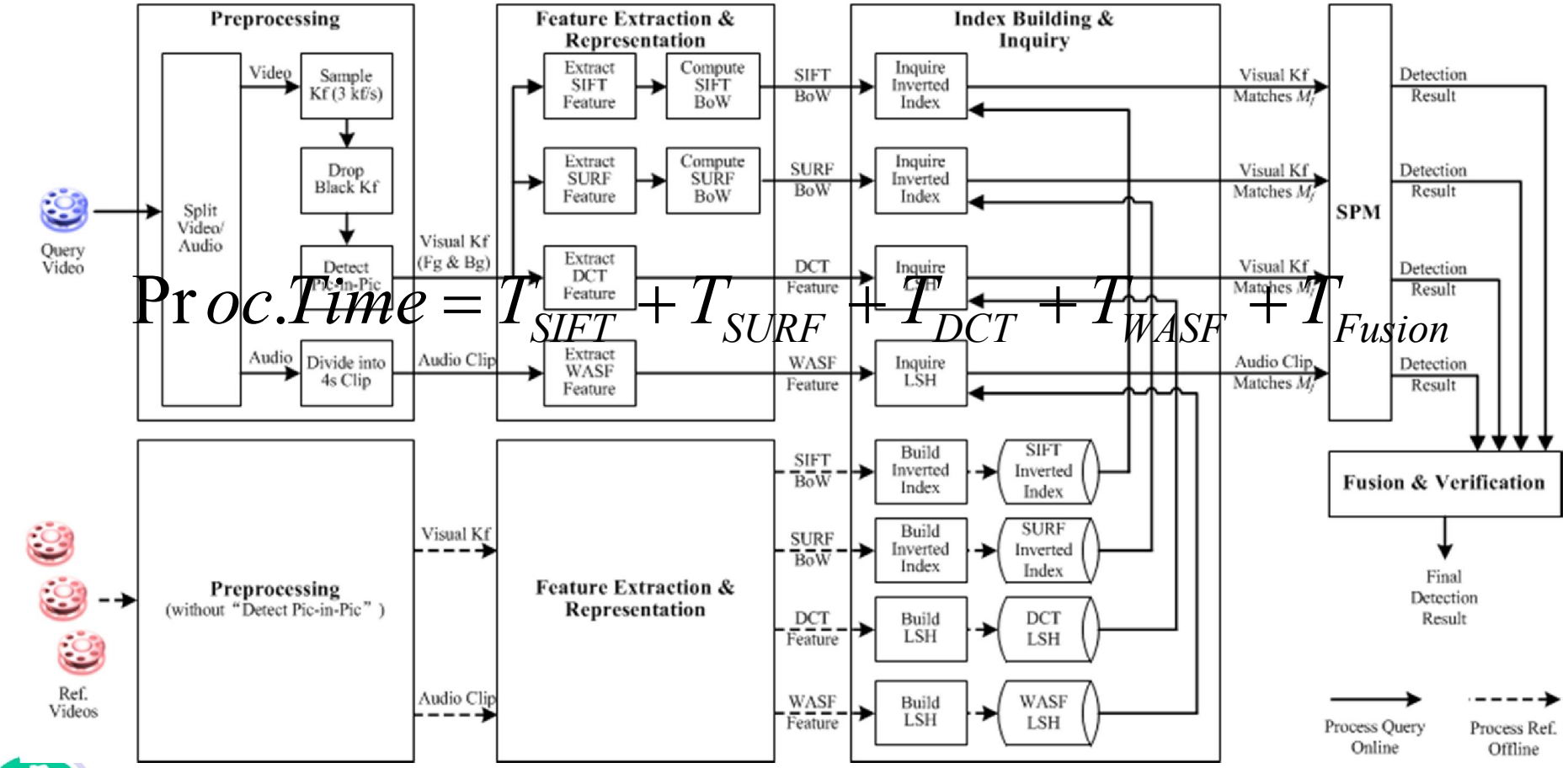
- Performance of DCSIFT detector with “TPM” vs. “HMM” on CCD10 and CCD09

Metrics	Methods	Dataset	V1	V2	V3	V4	V5	V6	V8	V10	AVG
NDCR	TPM	CCD10	0.285	0.154	0.054	0.146	0.038	0.223	0.292	0.200	**0.174**
		CCD09		0.112	0.030	0.090	0.024	0.142	0.201	0.149	**0.107**
	HMM	CCD10	0.346	0.207	0.131	0.200	0.116	0.285	0.354	0.269	0.239
		CCD09		0.164	0.090	0.142	0.090	0.194	0.245	0.187	0.159
M F1	TPM	CCD10	0.890	0.945	0.928	0.923	0.934	0.891	0.901	0.918	**0.916**
		CCD09		0.937	0.934	0.939	0.947	0.904	0.896	0.923	**0.926**
	HMM	CCD10	0.901	0.918	0.909	0.913	0.912	0.907	0.916	0.910	0.911
		CCD09		0.916	0.921	0.917	0.920	0.914	0.913	0.919	0.917
Time (s)	TPM	CCD10	0.004	0.004	0.004	0.004	0.004	0.004	0.004	0.004	**0.004**
		CCD09		0.004	0.004	0.004	0.004	0.004	0.004	0.004	**0.004**
	HMM	CCD10	0.103	0.102	0.103	0.103	0.103	0.103	0.103	0.103	0.103
		CCD09		0.102	0.101	0.101	0.102	0.102	0.103	0.101	0.102

S. K. Wei, et al., ”Frame fusion for video copy detection,” *IEEE TCSVT*, 21(1), 15–28, 2011.

(4) Cascade Architecture

- Our approach @ CCD10 – Late Fusion Strategy

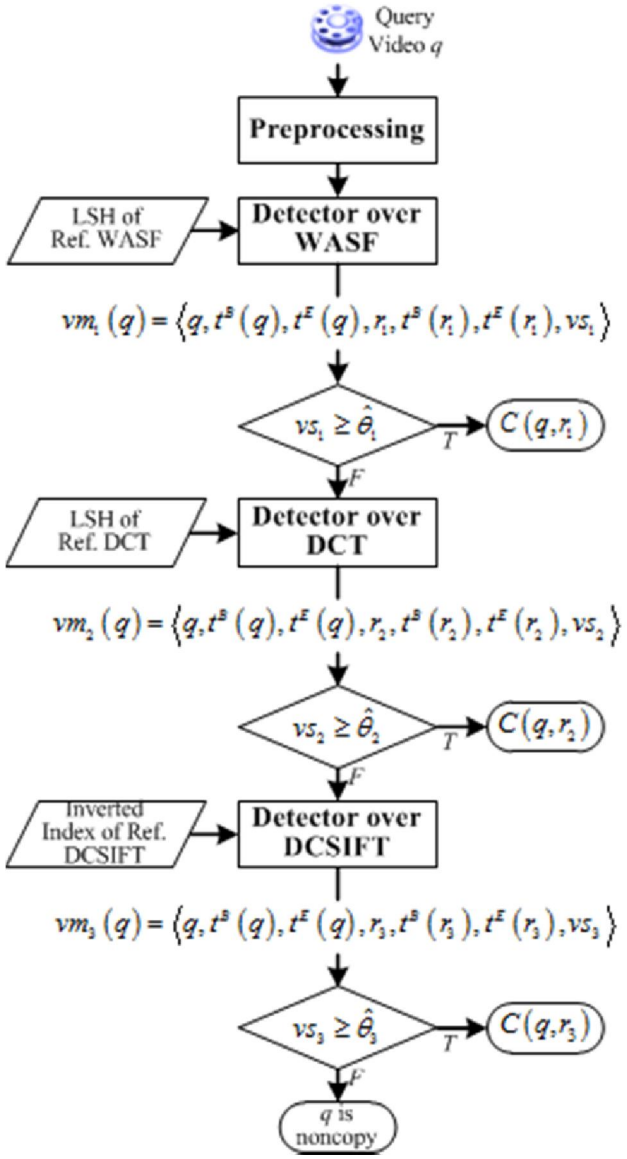


$$Proc.Time = T_{SIFT} + T_{SURF} + T_{DCT} + T_{WASF} + T_{Fusion}$$

Cascade Architecture

- Motivation
  - To be more efficient (compared with late fusion strategy)
  - To be more effective
- Design
  - Given a list of basic detectors
  - Place efficient yet ordinary detectors in the head
    - E.g., WASF, DCT
  - Put effective yet complex detectors in the tail
    - E.g., DCSIFT
- Task
  - *N*-Stage cascade $D_N = \langle d_1, d_2, \cdots, d_N \rangle$ with detectors $d_i, i = 1, 2, \cdots, N$
  - The problem: how to determine the decision thresholds

Cascade Architecture



$$
\begin{aligned}
&calculate\ vm_1(q)\\
&if\ (vs_1 \geq \theta_1)\\
&\quad return\ C(q, r_1)\\
&else\ \{\\
&\quad calculate\ vm_2(q)\\
&\quad if\ (vs_2 \geq \theta_2)\\
&\qquad return\ C(q, r_2)\\
&\quad else\ \{\\
&\qquad \ldots\\
&\qquad\quad calculate\ vm_N(q)\\
&\qquad\quad if\ (vs_N \geq \theta_N)\\
&\qquad\qquad return\ C(q, r_N)\\
&\qquad\quad else\\
&\qquad\qquad return\ NonCpy(q)\\
&\qquad \ldots\\
&\quad \}\\
&\}
\end{aligned}
$$

Parameters to be tuned:
Decision thresholds for all basic detectors $\{\theta_i\}_{i=1,2,\ldots,N}$

Where vm means video-level matches and vs means video-level similarity.

Cascade Architecture

- Enhance efficiency
  - Most copy queries are processed by WASF and DCT only!

<table>
<tr><th></th><th>A1</th><th>A2</th><th>A3</th><th>A4</th><th>A5</th><th>A6</th><th>A7</th></tr>
<tr><td>V1</td><td colspan="4" rowspan="7">Case1: WASF Only</td><td colspan="3">Case3: WASF+DCT+DCSIFT</td></tr>
<tr><td>V2</td></tr>
<tr><td>V3</td><td colspan="3" rowspan="5">Case2: WASF+DCT</td></tr>
<tr><td>V4</td></tr>
<tr><td>V5</td></tr>
<tr><td>V6</td></tr>
<tr><td>V8</td><td colspan="3" rowspan="2">Case3:WASF+DCT+DCSIFT</td></tr>
<tr><td>V10</td></tr>
</table>

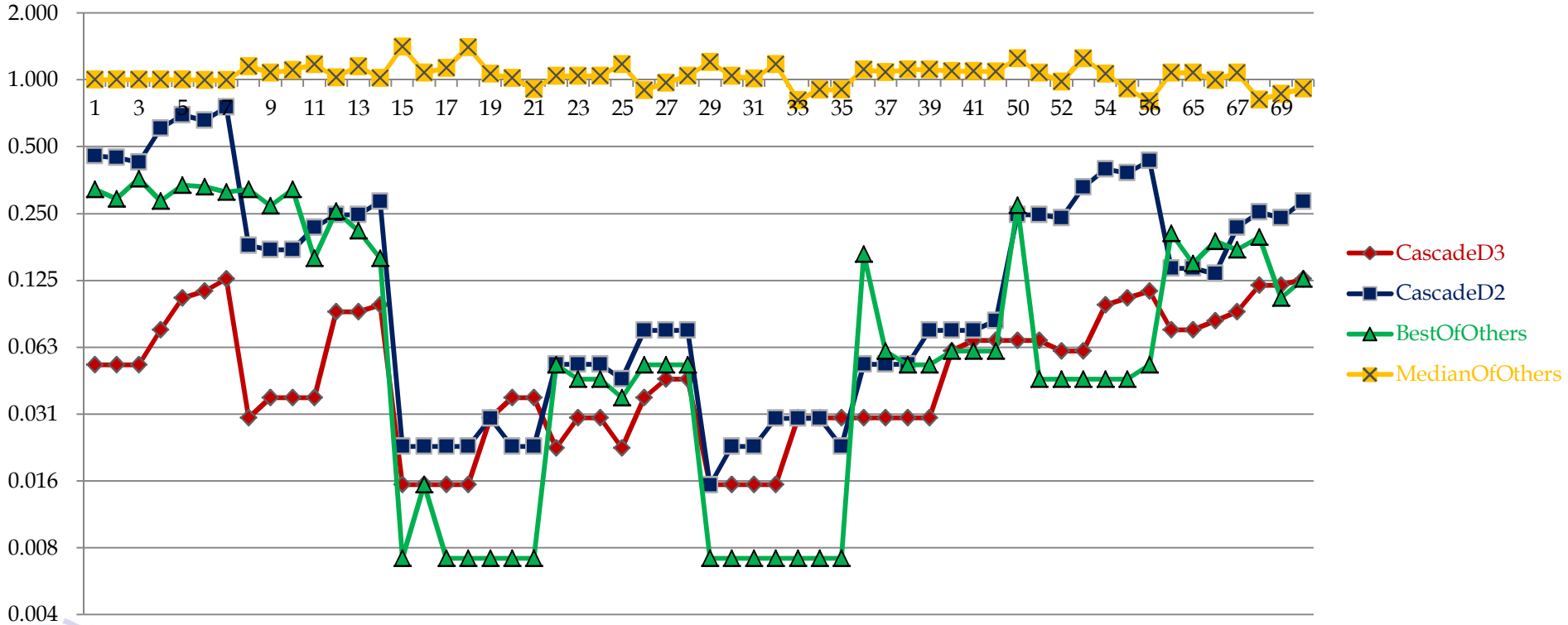
Evaluation Results

- Two approaches
  - CascadeD3: $D_3 = \langle d_{WASF}, d_{DCT}, d_{DCSIFT} \rangle$
  - CascadeD2: $D_2 = \langle d_{WASF}, d_{DCT} \rangle$
- Compelling performance ☺
  - Excellent NDCR
    - **34/56** best NDCR for BALANCED profile
    - **31/56** best NDCR for NOFA profile
  - Competitive MeanF1
    - **~0.95** for both profiles and all the transformations
  - Better-than-median/Almost-best MeanProcTime
    - CascadeD3: **172** sec/qry
    - CascadeD2: **11.75** sec/qry

All experiments are carried on an Windows Server 2008 with 32 Core 2.00 GHz CPUs and Memory-32 GB.

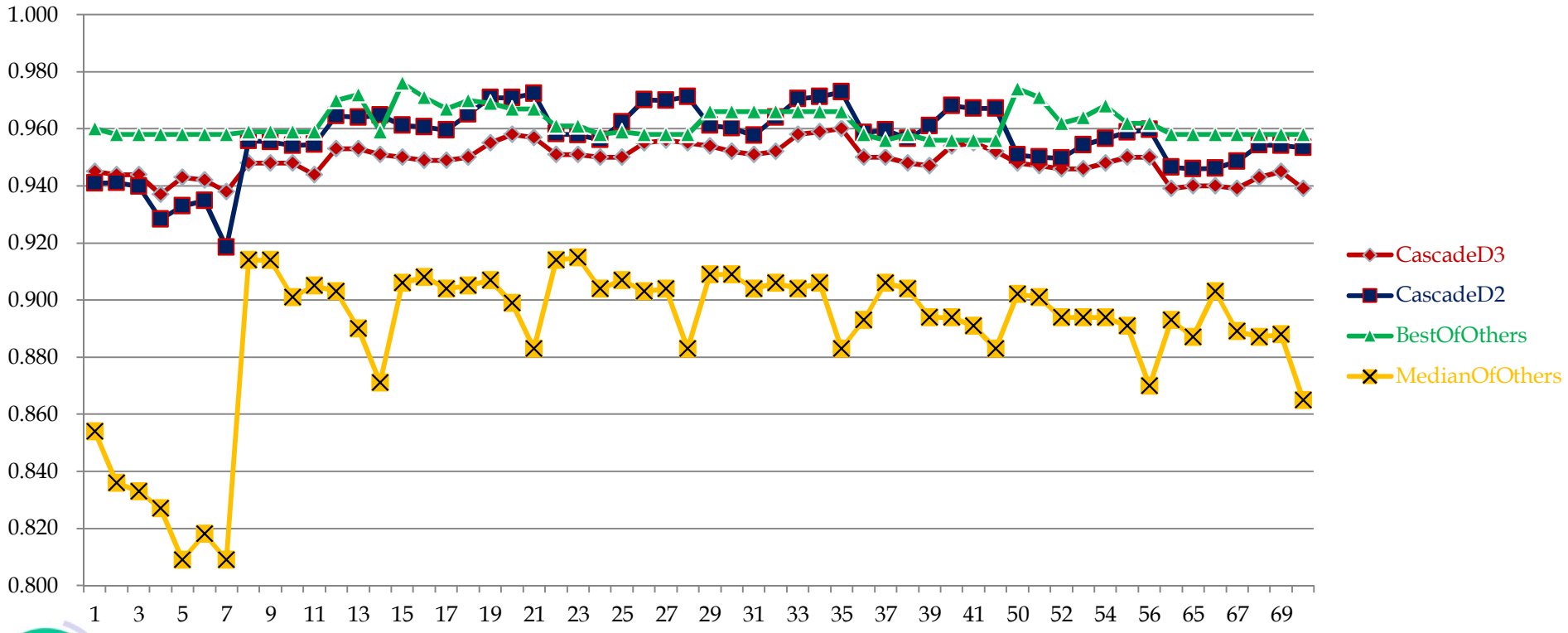
Evaluation Results

- Actual NDCR for BALANCED profile
  - CascadeD3: 34 best
  - CascadeD2: 12 outperform BestOfOthers



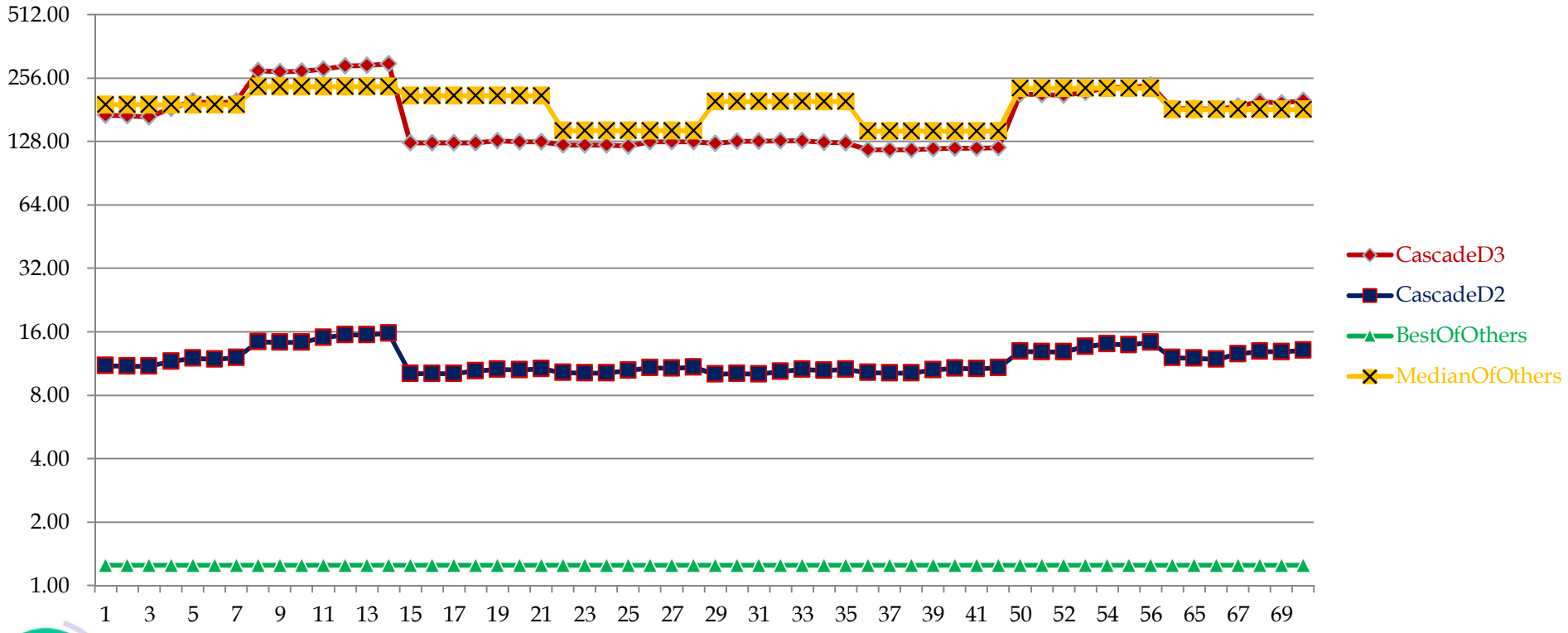
Evaluation Results

- Actual MeanF1 for BALANCED profile
  - CascadeD3: 0 best
  - CascadeD2: 17 best



Evaluation Results

- MeanProcTime for BALANCED profile
  - CascadeD3: 172 sec/qry
  - CascadeD2: 10.75 sec/qry



Recent Extension: Soft Cascade

- Above-mentioned Cascade Architecture
  - Employ hard (manually defined) decision thresholds
    $\Theta = \{\theta_1, \theta_2, \cdots, \theta_N\}$
  - **Hard Cascade!**

- Drawbacks of Hard Cascade architecture
  - Elaborately tuning of thresholds is burdensome
  - May not reach the optimal performance
  - Lack in generalization ability

Soft Cascade

- Soft Cascade Architecture
  - Learn the optimal decision thresholds (soft thresholds) $\hat{\Theta} = \{\hat{\theta}_1, \hat{\theta}_2, \cdots, \hat{\theta}_N\}$ automatically
- Key ideas
  - $\hat{\theta}_i$ should bring about a good tradeoff between FPs and FNs, and lead to the minimum error rate of d_i
  - The following detectors should focus on the queries which are incorrectly detected by previous detectors

M.-L. Jiang, Y.-H. Tian, T.-J. Huang, "Video Copy Detection Using a Soft Cascade of Multimodal Features," *IEEE ICME'12*, Under Review.

Soft Cascade

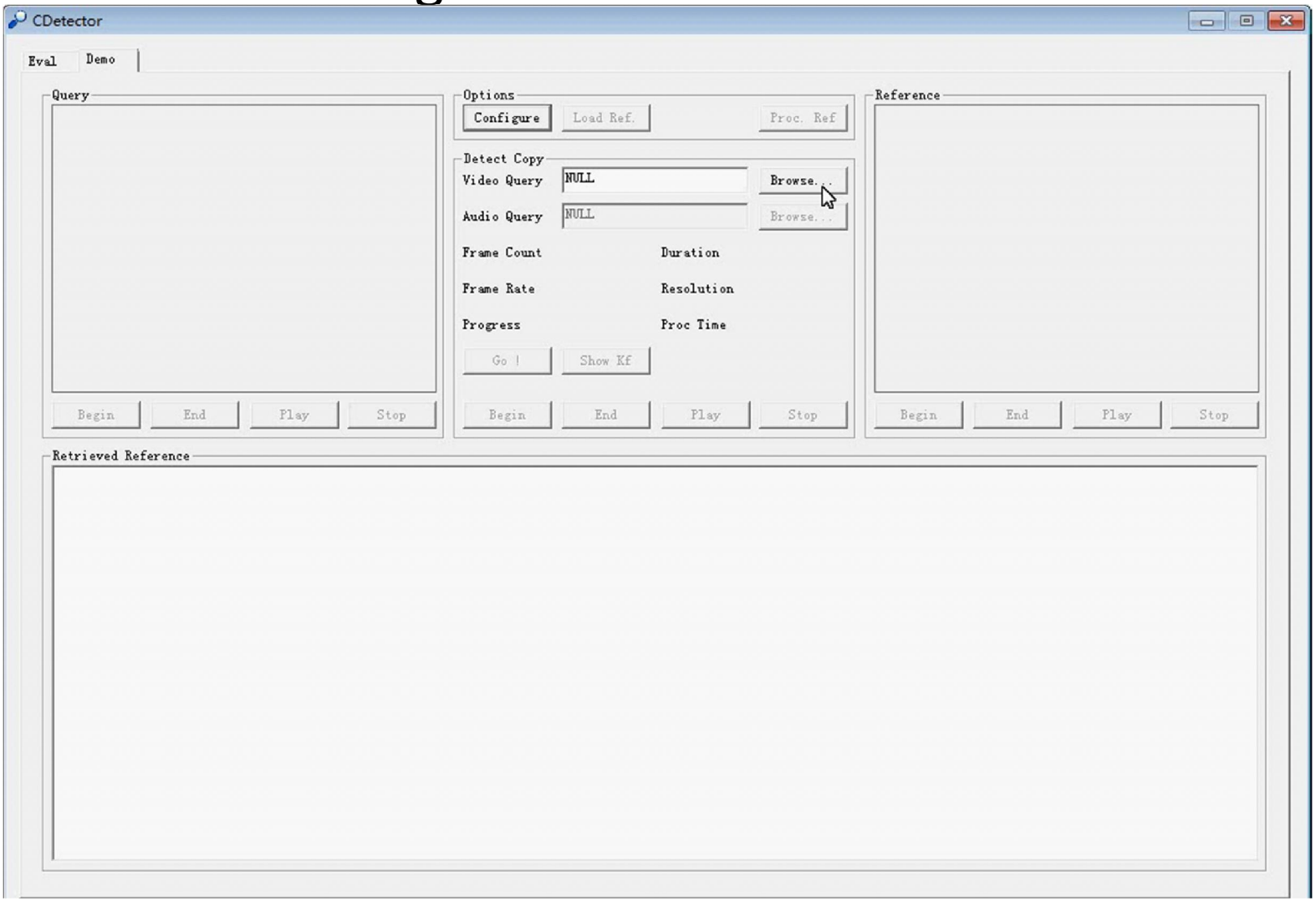
- Performance comparison between hard cascade, soft cascade and other participants' approaches

Approach		Avg. NDCR	Avg. MeanF1	Avg. MeanP.T.
Cascade Architecture	CascadeD3	0.060	0.951	172.291
	SoftD3	**0.054**	0.951	163.184
	CascadeD2	0.181	0.950	10.750
	SoftD2	0.178	0.950	**9.752**
Others	BestOfOthers	0.117	**0.962**	**1.250**
	MedianOfOthers	1.050	0.889	191.535

One of other participants' approaches could process a query within 1.30 seconds, but it suffers from high NDCR (Avg. NDCR=6.408) and low MeanF1 (Avg. MeanF1=0.001).

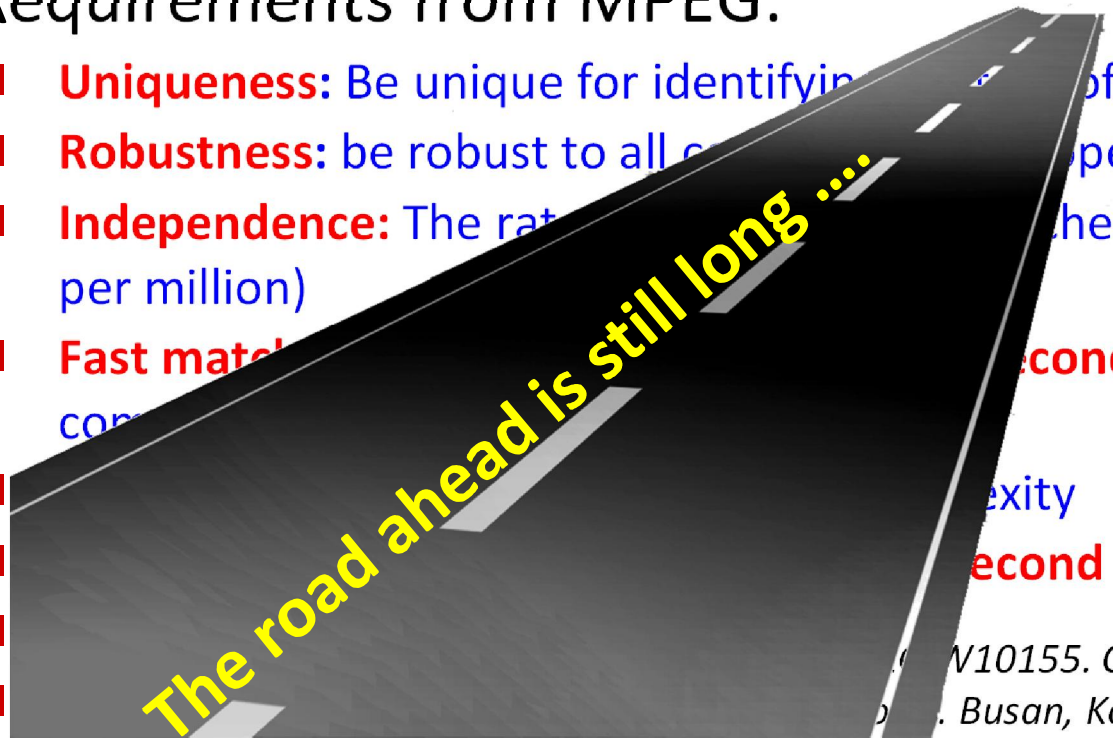
Demo

- CDetector Using CascadeD2



Summary: *CCD--- Ready for Application?*

- Video copy detection: A solved problem?
  - Best of our results: avg. NDCR= 0.054, Mean F1 = 0.951, Avg. Mean Processing Time < 3s
  - To further improve the performance: V3, V5, V8
- Requirements from MPEG:
  - **Uniqueness:** Be unique for identifyin… …f visual media
  - **Robustness:** be robust to all c… …perations
  - **Independence:** The rat… …hes ≤1 ppm (part per million)
  - **Fast matc**… …**cond** on a PC-class com…
  - …exity
  - …**econd** of content
  - …
  - …



…V10155. Call for proposals on video … Busan, Korea, Oct 2008.

THANKS

Member: Yonghong Tian, Menglin Jiang, Shu Fang
Tiejun Huang, Wen Gao

National Engineering Laboratory for Video Technology, Peking University