Hierarchical Hybrid Statistic based Video Binary Code and Its Application to Face Retrieval in TV-Series

Yan Li, Ruiping Wang, Shiguang Shan, Xilin Chen

Institute of Computing Technology,
Chinese Academy of Sciences

Background

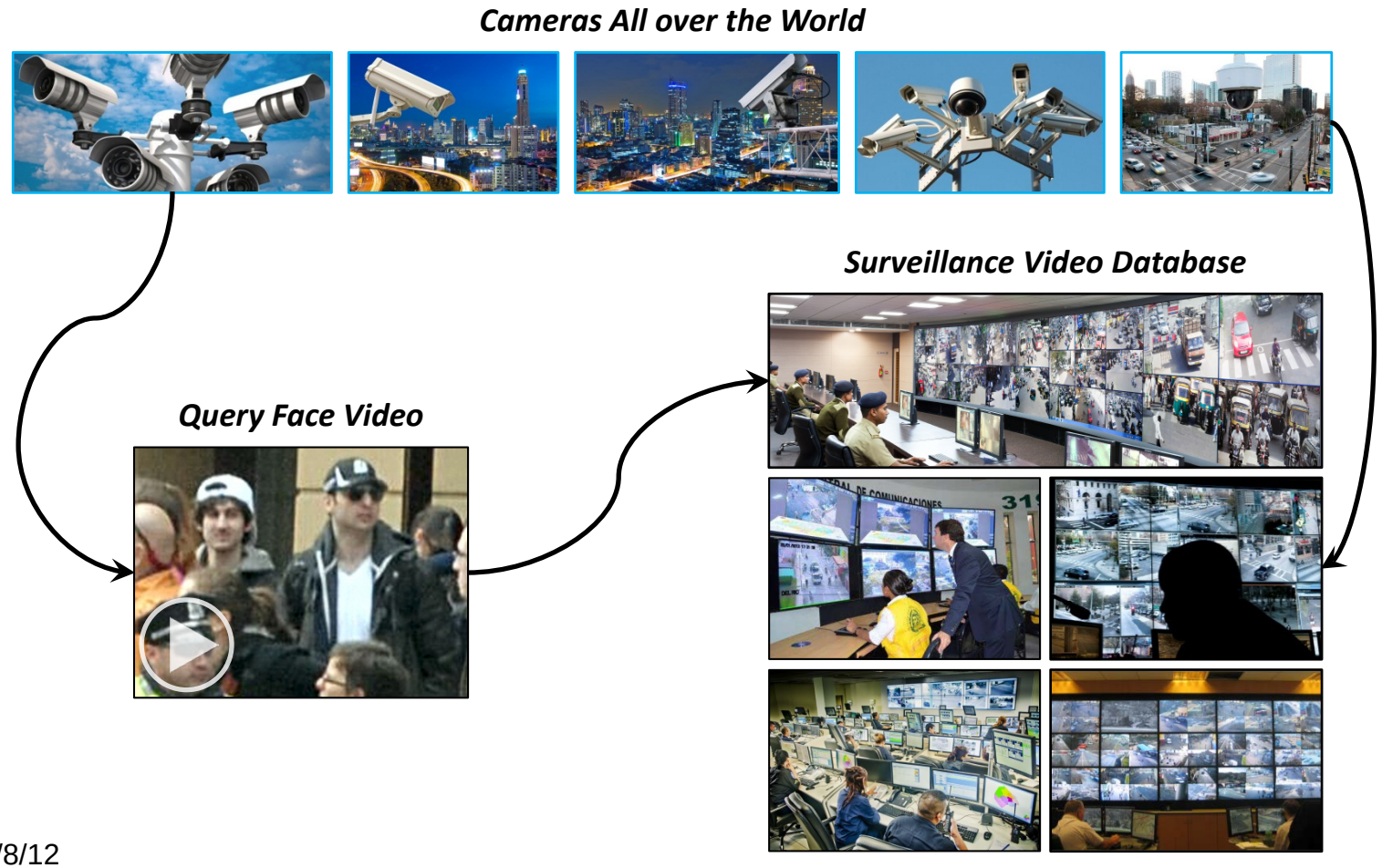
- Character Shots Retrieval from TV-Series / Movies

Query Face Video ***Database of Video Clips***

An example from "Prison Break"

Background

- Find a person from surveillance video

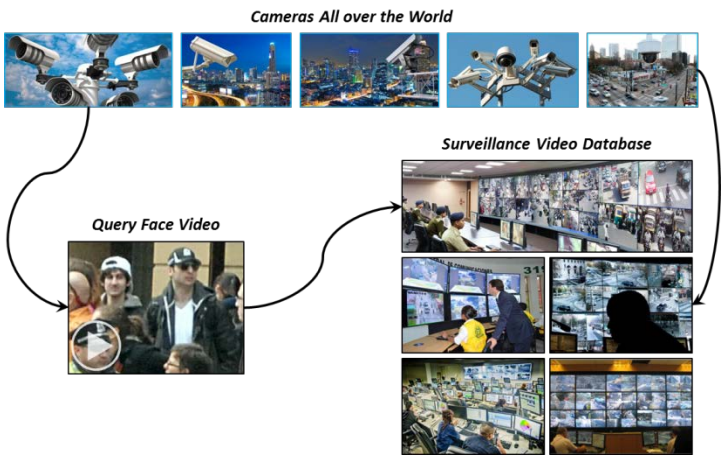


Problem

Q1: How to represent each face frame?

Q2: How to represent each face clip?

Q3: How to conduct retrieval efficiently?



Modeling a Face Frame

- Fisher Vector are utilized to model a face frame
  - Reasons
    - It is a Bag-of-Visual words (BoV) represent with soft assignment
    - It simultaneously encodes the zero, first and second-order statistic information

Modeling a Face Frame

- Fisher Vector are utilized to model a face frame
- Procedure of Fisher Vector Extraction
  - GMM Training
    - Step1: Local feature (dense SIFT) extraction from training data
    - Step2: Training GMM model

Modeling a Face Frame

- Fisher Vector is utilized to model a face frame
- Procedure of Fisher Vector Extraction
  - GMM Training
  - Modeling
    - Step1: Local feature (dense SIFT) extraction for a face image
    - Step2: 1st and 2nd statistic computation (for each feature dimension)

$$\Phi_k^{(1)} = \frac{1}{T\sqrt{w_k}} \sum_{t=1}^{T} \gamma_t(k) \left(\frac{x_t - \mu_k}{\sigma_k} \right)$$

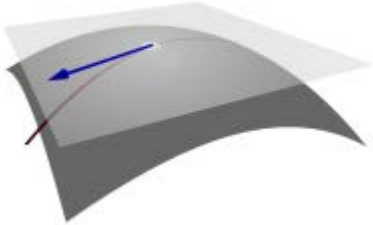
$$\Phi_k^{(2)} = \frac{1}{T\sqrt{w_k}} \sum_{t=1}^{T} \gamma_t(k) \frac{1}{\sqrt{2}} \left[\frac{(x_t - \mu_k)^2}{\sigma_k^2} - 1 \right]$$

    - Step3: Final Face Representation

$$\emptyset = \left[\Phi_1^{(1)}, \Phi_1^{(2)}, \ldots, \Phi_K^{(1)}, \Phi_K^{(2)} \right]$$

Modeling a Face Clip

- Sample Covariance Matrix is used for face clip modeling
- Sample Covariance Matrix



$$C = \frac{1}{n-1}\sum_{i=1}^{n}(f_i - \bar{f})(f_i - \bar{f})^T, C \in Sym_d^+$$

  - Riemannian Metric
  - Map to Reproducing Kernel Hilbert Space (RKHS) with Log-Euclidean Distance (LED)

$$d_{LED}(C_i, C_j) = \left\|\log(C_i) - \log(C_j)\right\|_F$$

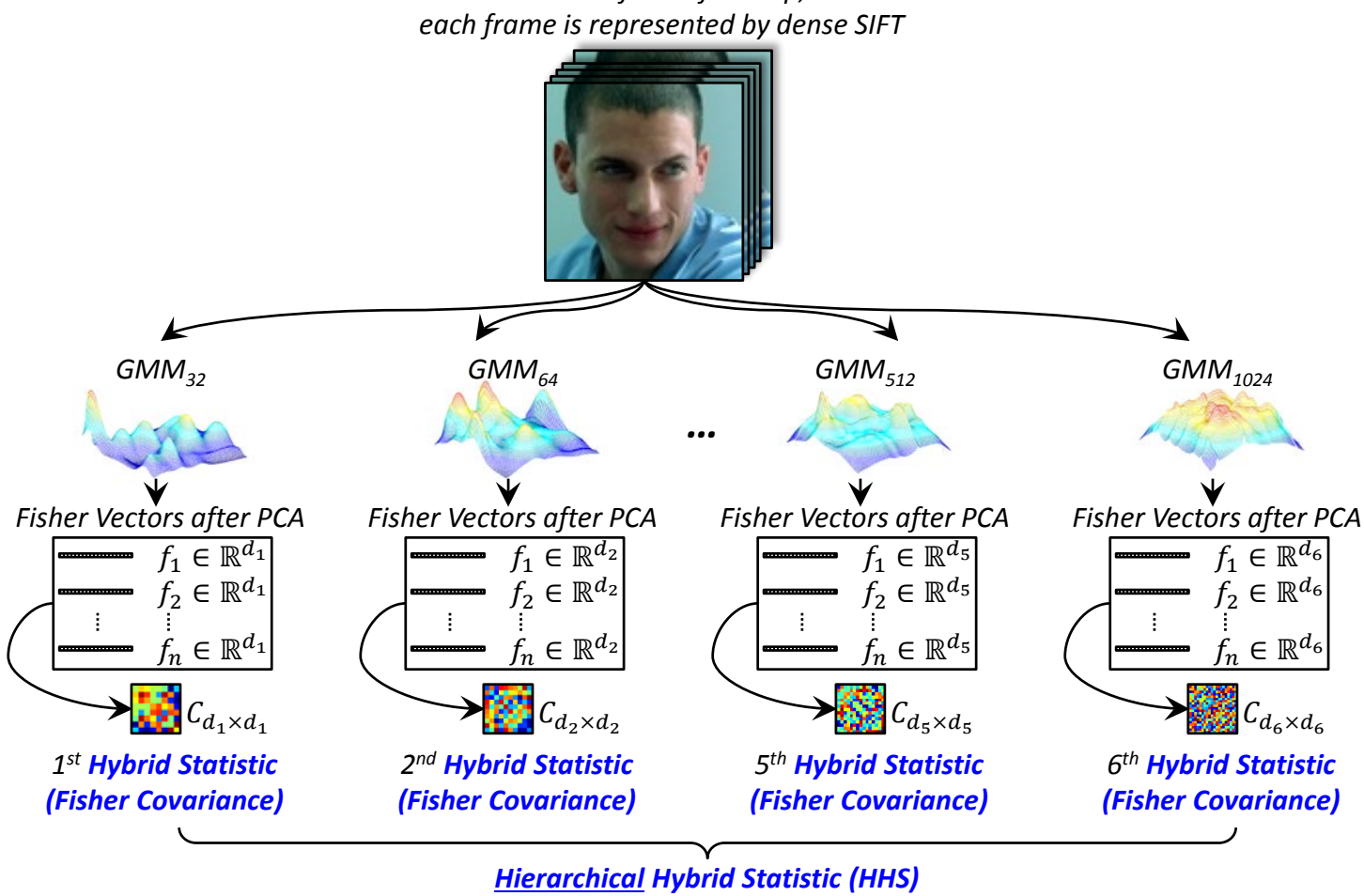
$$k_{\log}(C_i, C_j) = \text{trace}[\log(C_i) \cdot \log(C_j)]$$

  - Hybrid Statistic, when f_i is Fisher Vector, and we can also call it Fisher Covariance

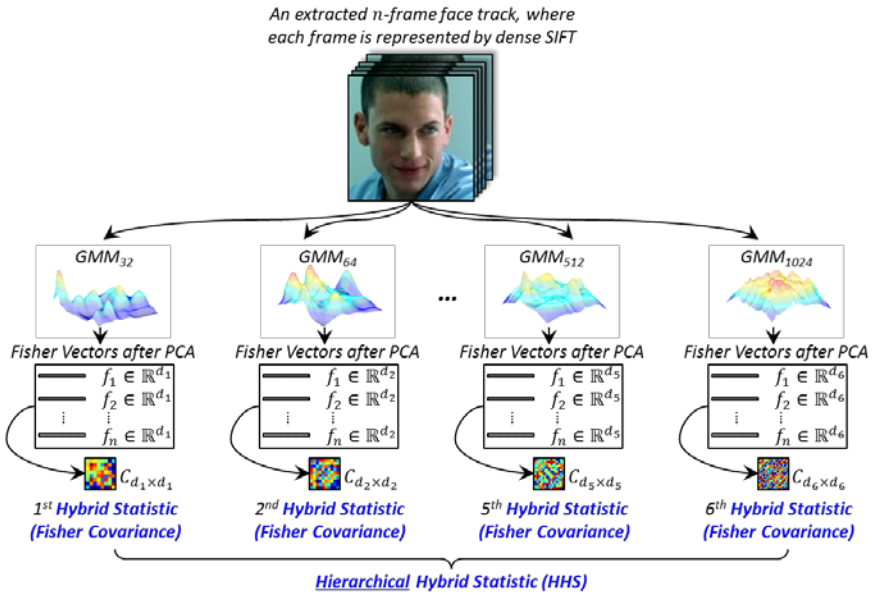
Modeling a Face Clip

- **K** in GMM for Fisher Vector

The Proposed Hierarchical Hybrid Statistic (HHS)



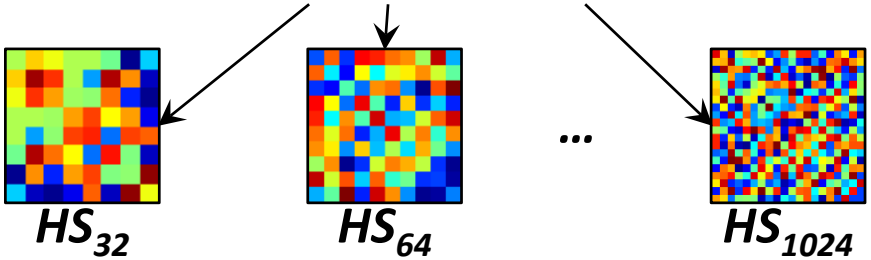
Modeling a Face Clip



What does HHS bring to us?

A coarse-to-fine face clip representation

$$C_i = \{C_{i1}, C_{i2}, \dots, C_{iH}\}$$



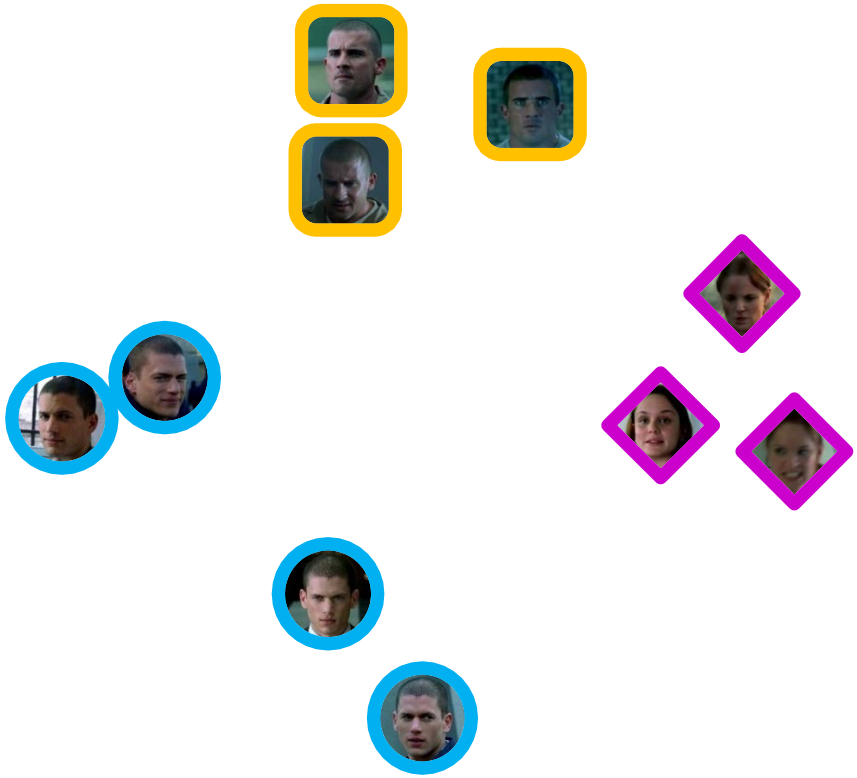
Binary Code Learning

- Challenge
  - HHS with high dimension
- Retrieval needs binary code, i.e., hash code.
- 32-bit codes can index more than 109 images ≈ the estimated number of images upload to Flickr in 2014



Binary Code Learning

Discriminability & Stability



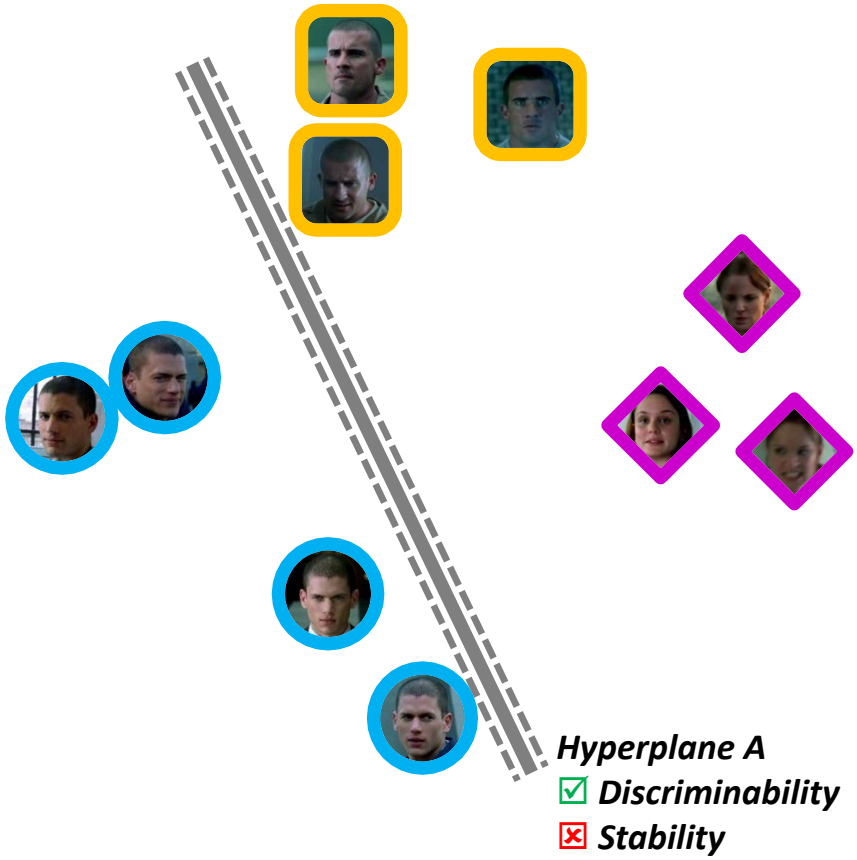
Two Properties

Discriminability

- *Intra-class compactness in Hamming space;*
- *Inter-class separability in Hamming space.*

Binary Code Learning

Discriminability & Stability



Two Properties

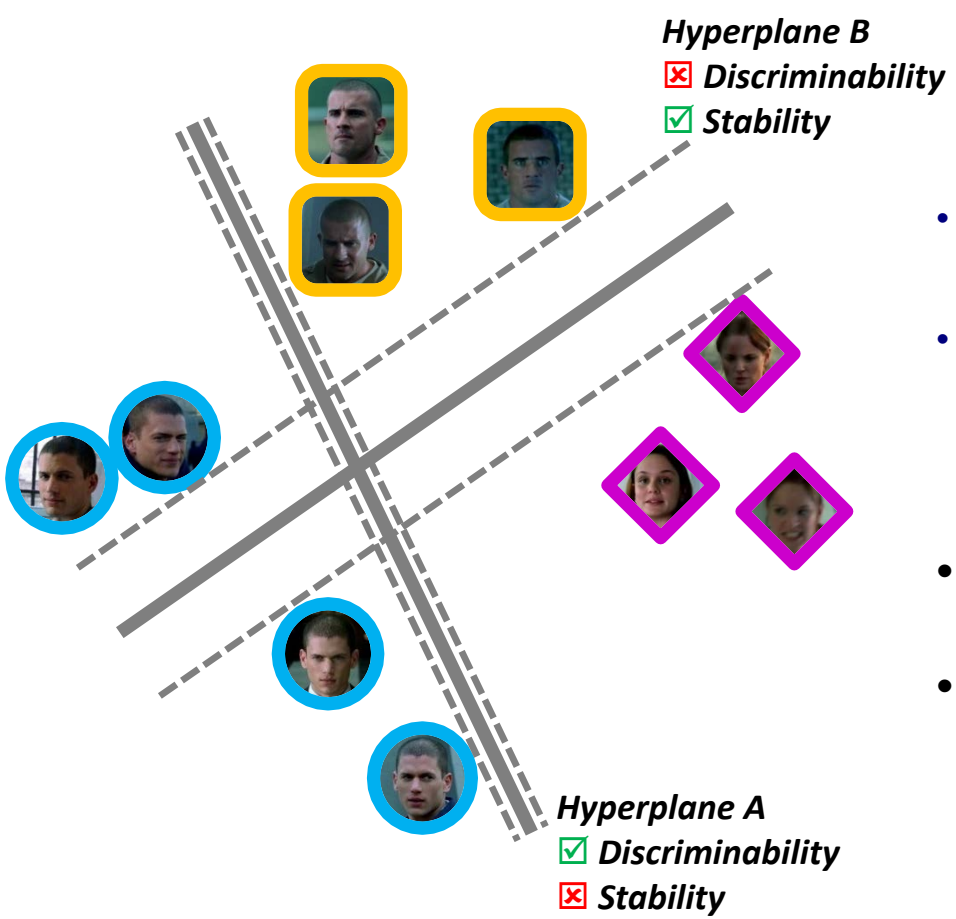
Discriminability

- *Intra-class compactness in Hamming space;*
- *Inter-class separability in Hamming space.*

Institute of Computing Technology, Chinese Academy of Sciences

Binary Code Learning

Discriminability & Stability



Two Properties

Discriminability

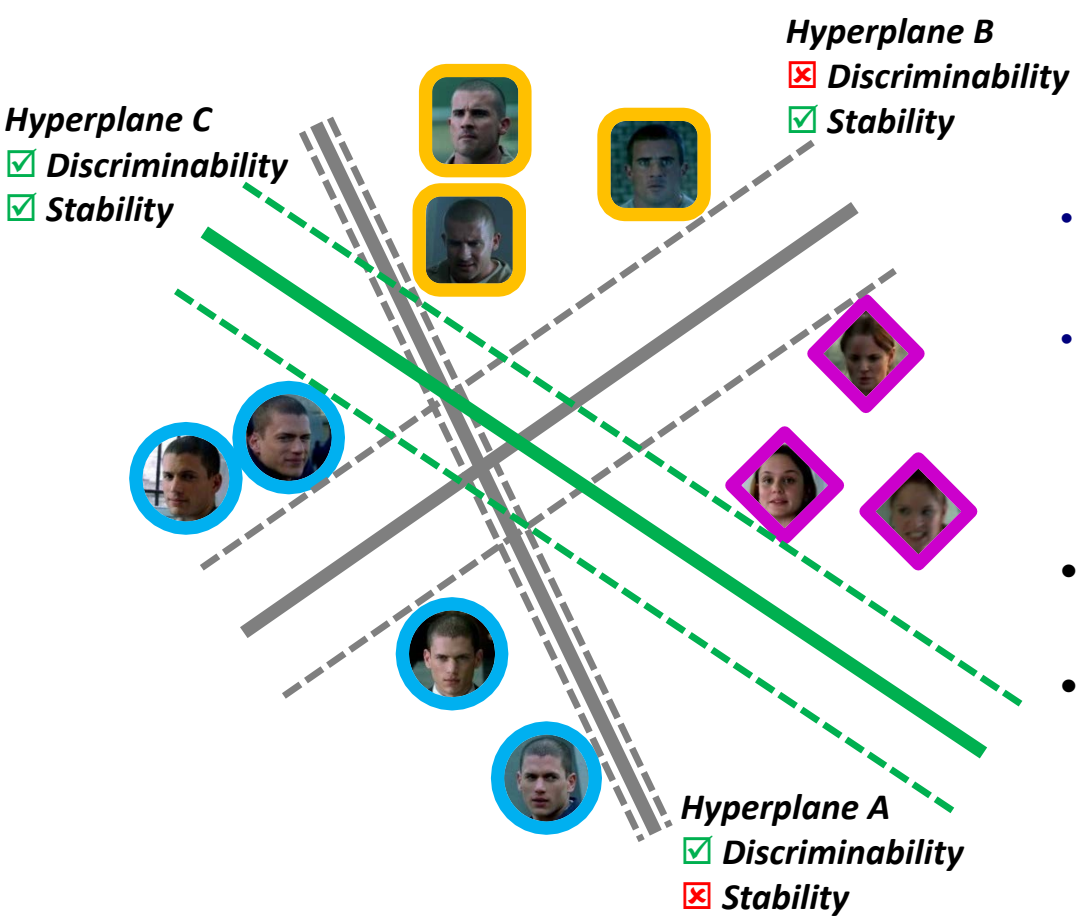
- *Intra-class compactness in Hamming space;*
- *Inter-class separability in Hamming space.*

Stability

- *Imagining each bit as a split in the original space;*
- *A split is stable when it has large margins from samples around it [Rastegari, ECCV'12].*

Binary Code Learning

Discriminability & Stability



Two Properties

Discriminability

- *Intra-class compactness in Hamming space;*
- *Inter-class separability in Hamming space.*

Stability

- *Imagining each bit as a split in the original space;*
- *A split is stable when it has large margins from samples around it [Rastegari, ECCV'12].*

Binary Code Learning

Iteratively Optimization

- *Objective Function*

$$\min_{\omega,\beta,b} \boldsymbol{E_{disc}} + \lambda \boldsymbol{E_{stab}}$$

- *Discriminability (LDA-like)*

$$\textit{Intra-class compactness: } S_W = \sum_{c\in\{1:R\}} \sum_{i,j\in c} dis(b_i, b_j)$$

$$\textit{Inter-class separability: } S_B = \sum_{\substack{c_1\in\{1:R\} \\ i\in c_1}} \sum_{\substack{c_2\in\{1:R\} \\ c_1\neq c_2, j\in c_2}} dis(b_i, b_j)$$

$$\textit{Energy function: } \boldsymbol{E_{disc}} = S_W - \lambda_1 S_B$$

- *Stability (SVM-like)*

$$\boldsymbol{E_{stab}} = \frac{1}{2}\|\omega\|^2 + \delta \sum_{i\in\{1:N\}} \max(1 - b_i \left(\omega^T \varphi(C_i)\right), 0)$$

Quick Summary

- Step 1. Face Clip Modeling
  - Frame Modeling: Fisher Vector
  - Clip Modeling: Sample Covariance Matrix
  - Hybrid Statistic: Fisher Covariance
  - Parameter selection: Hierarchical Hybrid Statistic
- Step 2. Binary Code Learning
  - Factor A: Discriminability
  - Factor B: Stability
  - Optimization: Block Co-ordinate Descent for iteratively optimizing

Experimental Dataset

- Two Hot TV-Series

the Big Bang Theory

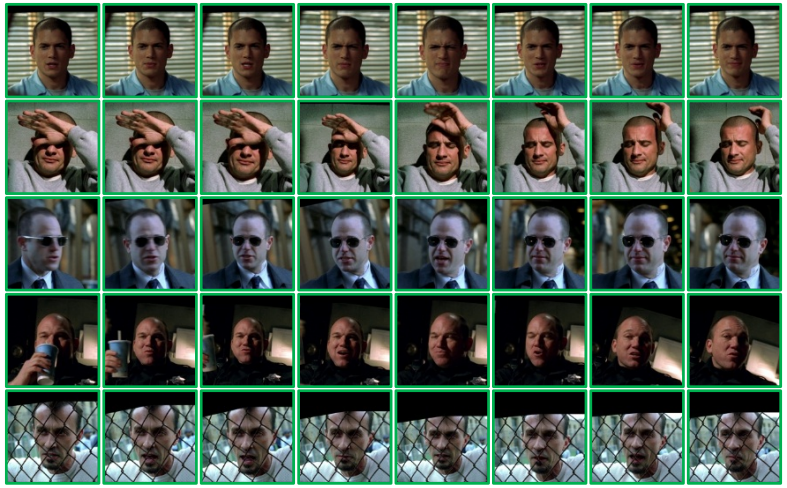
17 Episodes of Season 1

Prison Break

22 Episodes of Season 1

Experimental Dataset

- Key Tools
  1. Shot boundary detection;
  2. Face detection;
  3. Face tracking;
  4. Facial landmark localization;
  5. Character annotation (5 fans for each TV-Series).



Experimental Dataset

- Database Details
- Face Size: 80 X 64;
- Average Frame Num / clip: 45.
- Database Set
  - BBT: 4,527 videos of all the characters;
  - PB: 9,245 videos of all the characters.
- Test Protocol
  - Training Set (Randomly)
    - BBT: 140 videos of 14 characters;
    - PB: 190 videos of 19 characters.
  - Query Set
    - BBT: 50 videos of 5 main characters;
    - PB: 50 videos of 5 main characters.
- Measurements
  - mean Average Precision (mAP);
  - Precision Recall Curve.

Experimental Result

Feature	The Big Bang Theory					Prison Break				
	16 b	32 b	64 b	128 b	256 b	16 b	32 b	64 b	128 b	256 b
Gray + Cov	0.3510	0.3786	0.4032	0.4172	0.4430	0.1018	0.1042	0.1075	0.1135	0.1189
Gray (HE) + Cov	0.3662	0.4208	0.4666	0.4873	0.5120	0.0996	0.1054	0.1103	0.1134	0.1187
LBP + Cov	0.4653	0.5162	0.5332	0.5489	0.5678	0.1382	0.1507	0.1667	0.1801	0.2043
HOG + Cov	0.4874	0.5639	0.5998	0.6209	0.6479	0.1273	0.1464	0.1619	0.1726	0.1924
DSIFT + Cov	0.5319	0.6094	0.6453	0.6611	0.6803	0.1307	0.1518	0.1714	0.1837	0.2008
HS_{32}	0.6524	0.7325	0.7505	0.7774	0.7961	0.1458	0.1758	0.2069	0.2270	0.2616
HS_{64}	0.6663	0.7435	0.7731	0.8029	0.8131	0.1533	0.1853	0.2130	0.2321	0.2598
HS_{128}	0.7186	0.7829	0.8071	0.8257	0.8406	0.1519	0.1857	0.2167	0.2323	0.2577
HS_{256}	0.7213	0.7909	0.8195	0.8426	0.8600	0.1547	0.1822	0.2012	0.2189	0.2434
HS_{512}	0.7918	0.8494	0.8573	0.8663	0.8730	0.1552	0.1824	0.1996	0.2158	0.2365
HS_{1024}	0.8078	0.8597	0.8660	0.8731	0.8779	0.1608	0.1829	0.2009	0.2119	0.2303
HHS	**0.8763**	**0.9113**	**0.9078**	**0.9116**	**0.9172**	**0.1950**	**0.2279**	**0.2585**	**0.2743**	**0.3035**

- *Fisher Vector performs better than the other front-end face frame representations;*
- *Hierarchical structure further boosts the performance.*

Experimental Result

Comparing on different binary code learning methods

Feature	The Big Bang Theory					Prison Break				
	16 bits	32 bits	64 bits	128 bits	256 bits	16 bits	32 bits	64 bits	128 bits	256 bits
HS+LSH [Gionis, VLDB'99]	0.3783	0.4093	0.4148	0.4414	0.4383	0.0998	0.1048	0.1081	0.1078	0.1101
HS+RR [Gong, CVPR'11]	0.4207	0.4042	0.4507	0.4622	0.4407	0.0981	0.1018	0.1042	0.1065	0.1105
HS+ITQ [Gong, CVPR'11]	0.3445	0.4033	0.4257	0.4428	0.4324	0.1129	0.1083	0.1114	0.1095	0.1098
HS+SH [Weiss, NIPS'08]	0.4225	0.3802	0.3809	0.3765	0.3972	0.0901	0.0978	0.0964	0.1048	0.1059
HS+SSH [Wang, CVPR'10]	0.3134	0.2830	0.2757	0.2878	0.3656	0.1527	0.1488	0.1417	0.1409	0.1436
HS+KSH [Liu, CVPR'12]	0.5799	0.6506	0.6965	0.7094	0.7300	0.1571	0.1546	0.1619	0.1630	0.1599
HS+SITQ [Gong, CVPR'11]	0.6185	0.6702	0.6891	0.7006	0.7165	0.1211	0.1326	0.1462	0.1578	0.1640
HS_{512}	**0.7918**	**0.8494**	**0.8573**	**0.8663**	**0.8730**	**0.1552**	**0.1824**	**0.1996**	**0.2158**	**0.2365**
HHS	**0.8763**	**0.9113**	**0.9078**	**0.9116**	**0.9172**	**0.1950**	**0.2279**	**0.2585**	**0.2743**	**0.3035**

Institute of Computing Technology, Chinese Academy of Sciences

Conclusions

- Hybrid Statistic (Fisher Covariance) for face clip modeling
- Extend Hybrid Statistic (HS) to Hierarchical Hybrid Statistic (HHS)
- LDA-like discriminability and SVM-like stability constraints are used for binary code learning
- A database for face retrieval vipl.ict.ac.cn/resources/datasets.

TV-Series

Demo Show

Institute of Computing Technology, Chinese Academy of Sciences

THANKS

Institute of Computing Technology, Chinese Academy of Sciences