The FG 2015 Video Person Recognition Evaluation

Ross Beveridge[1], Hao Zhang[1], Bruce Draper[1], Patrick Flynn[2], Zhenhua Feng[3], Patrik Huber[3], Josef Kittler[3], Zhiwu Huang[4], Shaoxin Li[4], Yan Li[4], Meina Kan[4], Ruiping Wang[4], Shiguang Shan[4], Xilin Chen[4], Haoxiang Li[5], Gang Hua[5], Vitomir Struc[6], Janez Krizaj[6], Changxing Ding[7], Dacheng Tao[7] and Jonathon Phillips[8]

[1]Colorado State University, USA, [2]University of Notre Dame, USA, [3]University of Surrey, UK, [4]Institute of Computing Technology CAS , China, [5]Stevens Institute of Technology, USA, [6]University of Ljubljana, Slovenia, [7]University of Technology, Sydney, Australia, [8]National Institute for Standards and Technology, USA

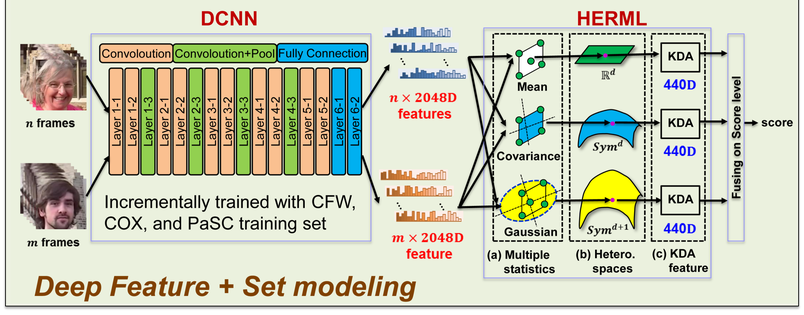
Overview

- Point-and-Shoot Face Recognition Challenge (PaSC) was introduced in 2013.
- At IJCB 2014, Handeld Video Face and Person Recognition Competition was presented.
- The FG 2015 Video Person Recognition Evaluation was carried out in the Fall of 2014.

Approach Summaries for the Five Participants

Ins. of Comp. Tech. CAS (ICT CAS)

- **Method**: **H**ybrid **E**uclidean-and-**R**iemannian **M**etric **L**earning with **De**eply **L**earned **F**eatures (**HERML-DeLF**)
- **Step 1**: DCNN for single frame feature learning/extraction
  - DCNN architecture:14 convolution layers, 2 fully connected layers and 1 softmax layer
  - DeLF: outputs of Layer 6-2（totally 2,048D）for each frame
- **Step 2**: HERML for image set modeling (integrating DCNN features of all frames)
  - Fusing three set-based video features : mean, covariance matrix and Gaussian model
  - Dimension of final KDA feature for each video: 1,320D(=3*440D)



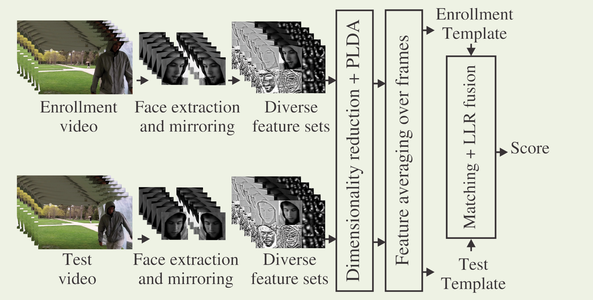
Deep Feature + Set modeling

University of Ljubljana (Uni-Lj)

University of *Ljubljana* Faculty *of Electrical Engineering*

MODEST framework

- Probabilistic **mo**deling of **d**iverse **f**eature **se**ts – MODEST
- Supplied eye coordinates → geometric normalization
- Luminance-reflectance model → photometric normalization
- Diverse feature extracted from images (Gabor, Pixels, LBP, LPQ)
- PLDA model trained for each feature set
- Score-level combination with Linear Logistic Regression (LLR)

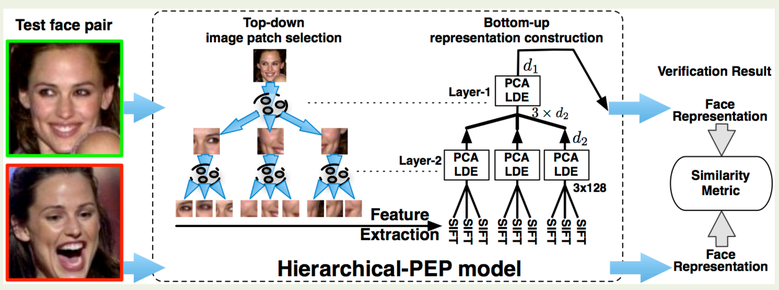


Geometric normalization crucial for performance

Stevens Institute of Technology (SIT)

Hierarchical-PEP Model

- Unified face representation for face image and face video
- Decompose face into parts of parts
- Hierarchically integrate parts with a PCA-LDE net
- Construct low dimensional pose invariant face representation

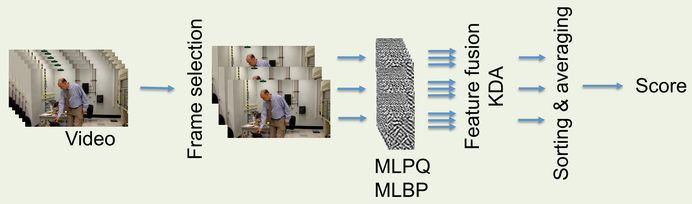


Fine-grained structures of faces parts help in addressing pose variations.

University of Surrey (Surrey)

MLPQ face matching on videos

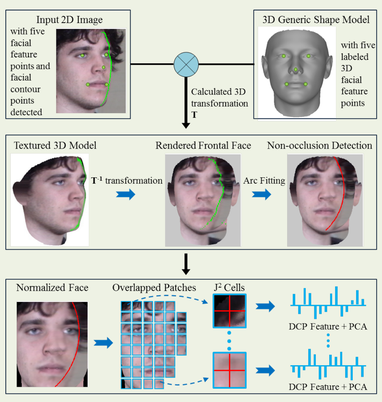
- Frame selection based on sharpness, focus, pose and face size
- Multiscale Local Phase Quantisation and LBP features are extracted and combined using kernel fusion
- Kernel Discriminant Analysis (KDA) of the combined features extracts discriminative information for face recognition
- The 20 highest frame scores are averaged



The "off the shelf" face matching algorithm outperforms LRPCA but it is challenged by the harsh conditions of PaSC images

University of Technology, Sydney (UTS)

Modified from the PBPR approach [1]



[1] C. Ding, C. Xu, and D. Tao, "Multi-task pose-invariant face recognition," IEEE TIP, 2015.

[2] C. Ding, D. Tao, "A comprehensive survey on pose-invariant face recognition," arXiv preprint, 2015.

<1> The PBPR approach handles the full range of pose variation from -90 degree to +90 degree and achieves strong performance. (please refer to [1] and [2])
<2> The modified approach from PBPR achieves very competitive performance with only limited training data in this competition.

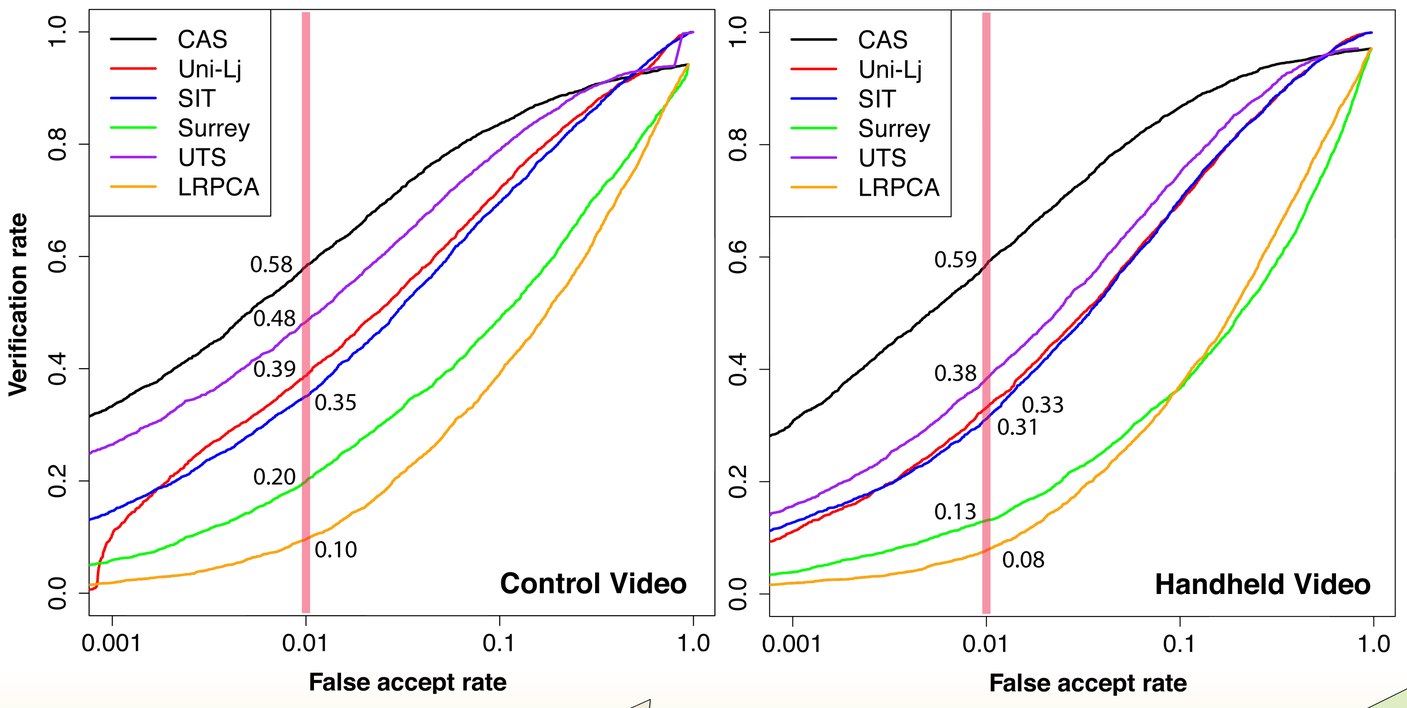
Performance Summary ROCs for the Evaluation

High Quality Video Camera - Control

- Tripod Mounted Panasonic HD700 camera
- Frame size is 1920x1080
- Total of 1401 videos of 265 people
- Participants delivered 1401^2 similarity matrices

Handheld Consumer Grade Cameras

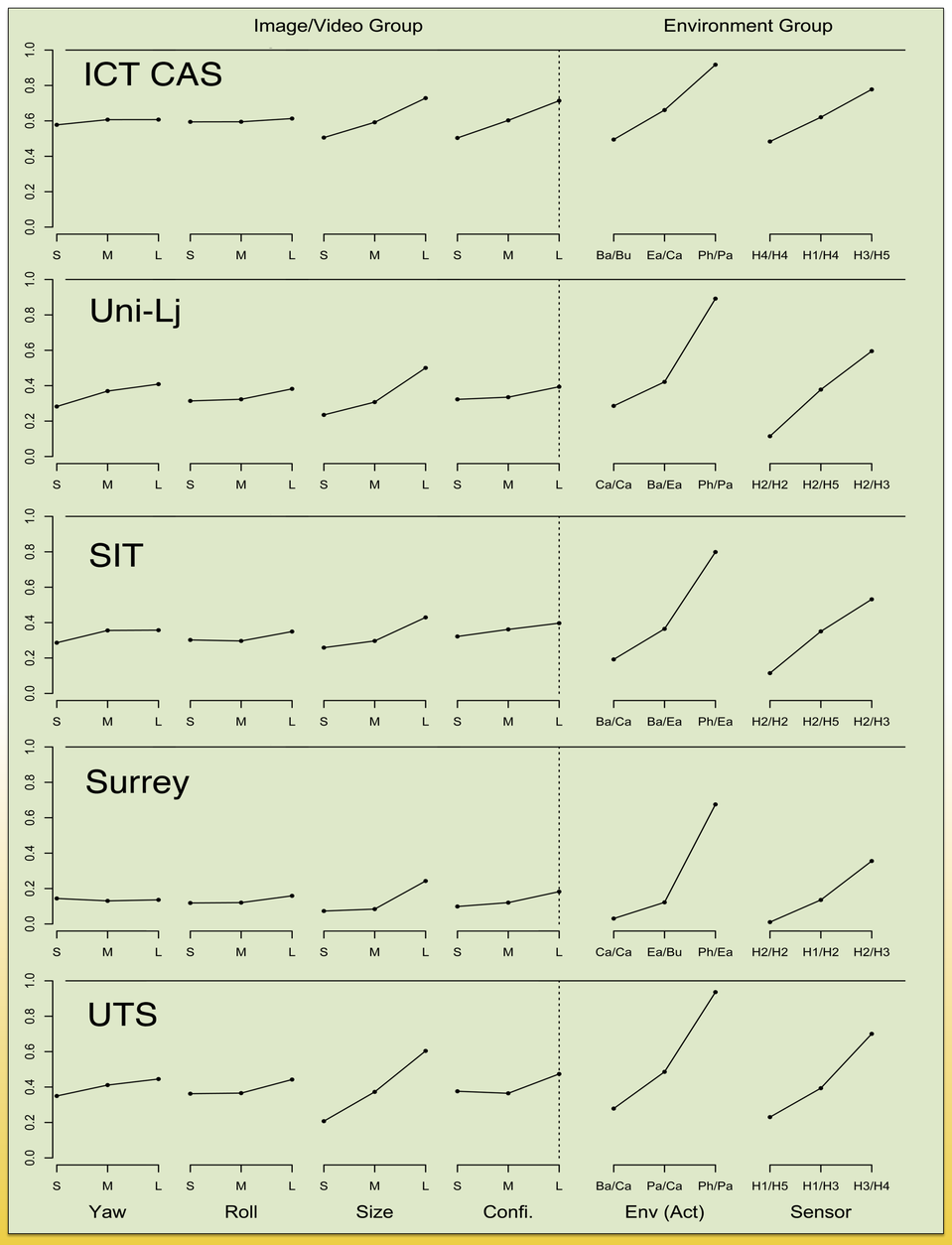
- Videos from a mix of 5 different cameras
- Camera types/sizes are summarized to the right
- Total of 1401 videos of 265 people
- Participants delivered 1401^2 similarity matrices



Covariates

Sensor
Size
Location
Action

Sensor	Size	Location	Action
Flip Mino F360B	640x480	Canopy	Golf Swing
Kodak Zi8	1280x720	Canopy	Bag Toss
Samsung M. CAM	1280x720	Office	Newspaper
Sanyo Xacti	1280x720	Lab 1	Write on Easel
Sanyo Xacti	1280x720	Lawn	Blow Bubbles
Nexus Phone	720x480	Stone	Ball Toss
Kodak Zi8	1280x730	Lab 2	Pickup Phone



Evaluation Performance Summary

- Wide range of approaches tested
- Wide range of performance, all better than LRPCA
- Significant Progress, best seen so far in ...
  - Either IJCB 2014 competition or the PittPatt SDK
- Control videos easier for most participants
- No Control vs. Handheld difference for top algorithm
- Still a lot of headroom for future improvements

Covariates Explanation & Conclusions

- Vertical axes is verification rate @ FAR = 0.01
- Plots highlight performance change
- Algorithms trend together
- Location/Action matter the most
- Camera matters, but is confounded with location
- Image/video attributes are secondary
- Among image/video attributes, size matters most