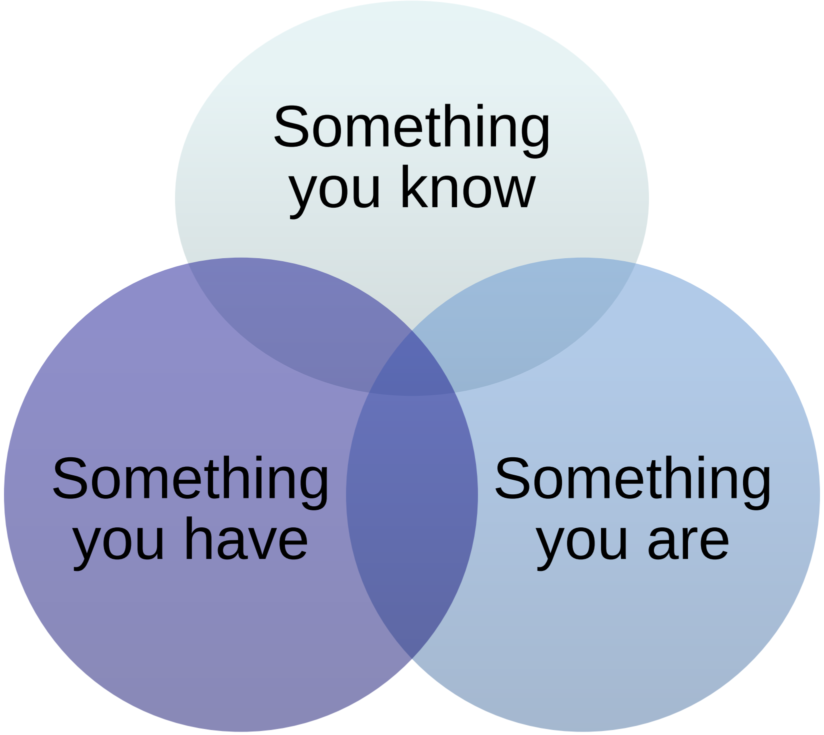
ROBUST FACE RECOGNITION for UNCONTROLLED SETTINGS

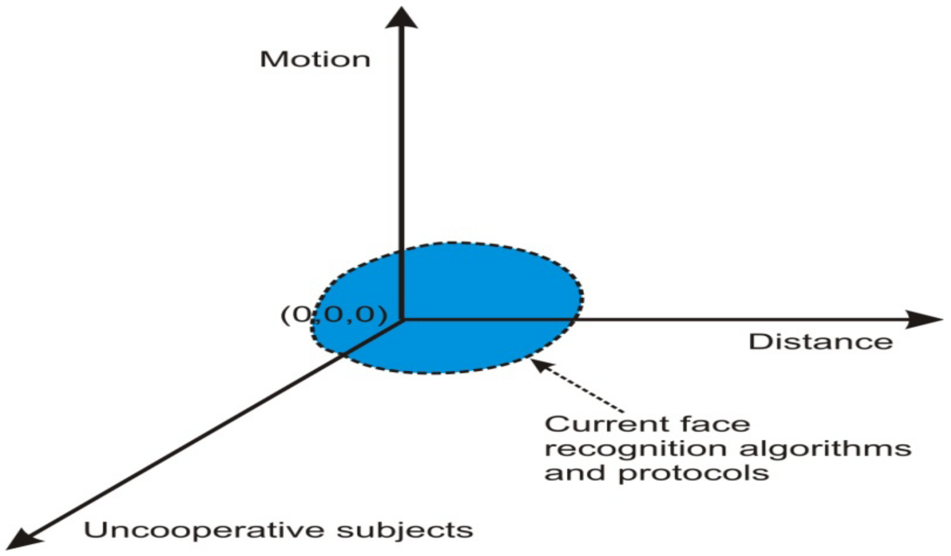
Harry Wechsler
Department of Computer Science
George Mason University
Fairfax, VA 22030
wechsler@gmu.edu

Personal Authentication Using Smart Identity Management

Knowledge, Token, and Biometrics // Appearance, Behavior, and Cognitive State //



Uncontrolled Settings and Labeled Faces in the Wild (LFW)



Terminology

Recognition ~ Categorization ~ Authentication

Authentication = {Verification, Identification, Surveillance / Watch List}

Categorization1 = {(Pedestrian) Detection: *Face in a Crowd* (Human, Face), *Stratification* / Binning: <Gender, Ethnicity, Age> ~ *Index1*, *Identification / Authentication*}

Categorization2 = Soft Biometrics ~ *Index2*

Categorization3 = Context / Situation Awareness, Knowledge & Logistics / *W5+* ~ *Index3 / Who, Where, What, Why, When, How /*

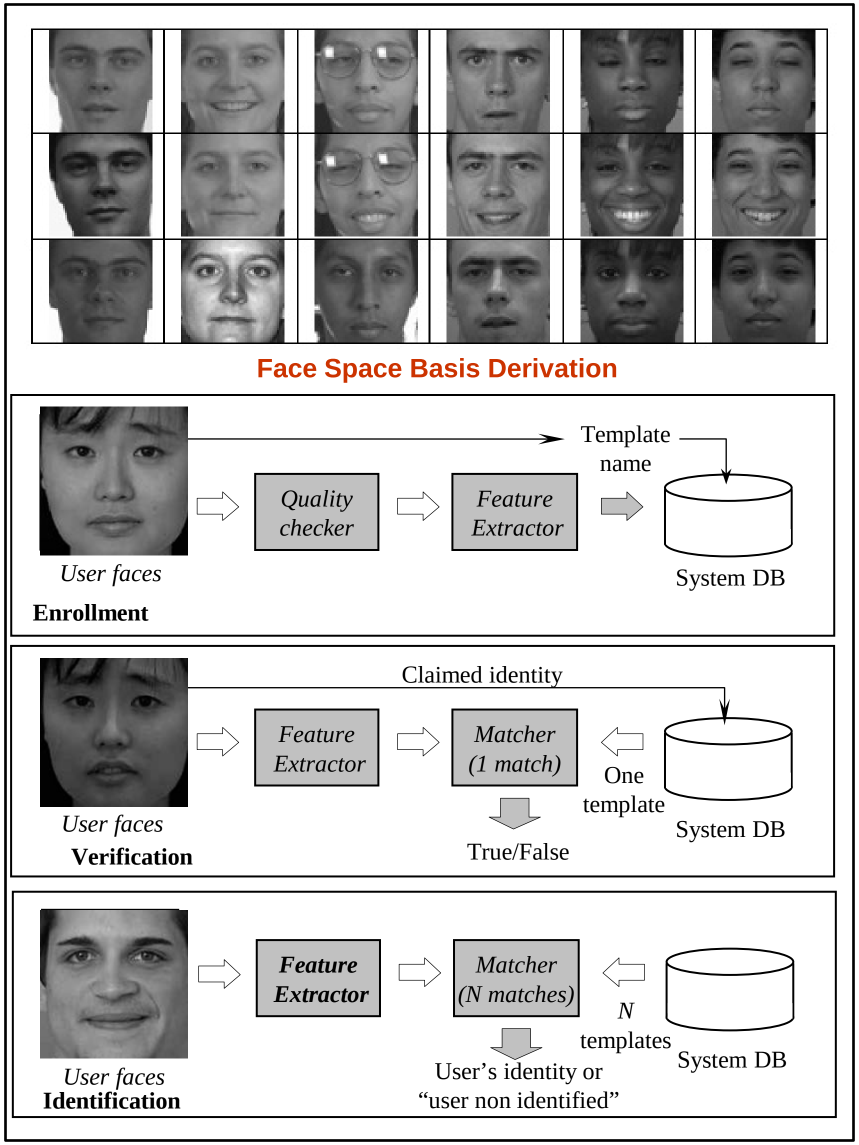
Applications

- Security, Mass Screening, and (BIG) Data Mining
- Authentication and Re-Identification / Face in a Crowd
- Photo Tagging and Social Networks
- Mobile and Sensor / Camera Networks for Surveillance
- Smart Biometric Spaces / Marketing and Retail /
- Wearable Devices
- Health Care (Assess / Monitor / Rehabilitate)
- Massive Open On-Line Courses (MOOC)

Identity Management

"I'm sorry, but someone else with that identity is already here."

Authentication Protocols



Face Space Basis Derivation

User faces

Quality checker

Feature Extractor

Template name

System DB

Enrollment

Claimed identity

User faces

Feature Extractor

Matcher (1 match)

One template

System DB

True/False

Verification

User faces

Feature Extractor

Matcher (N matches)

N templates

System DB

User's identity or "user non identified"

Identification

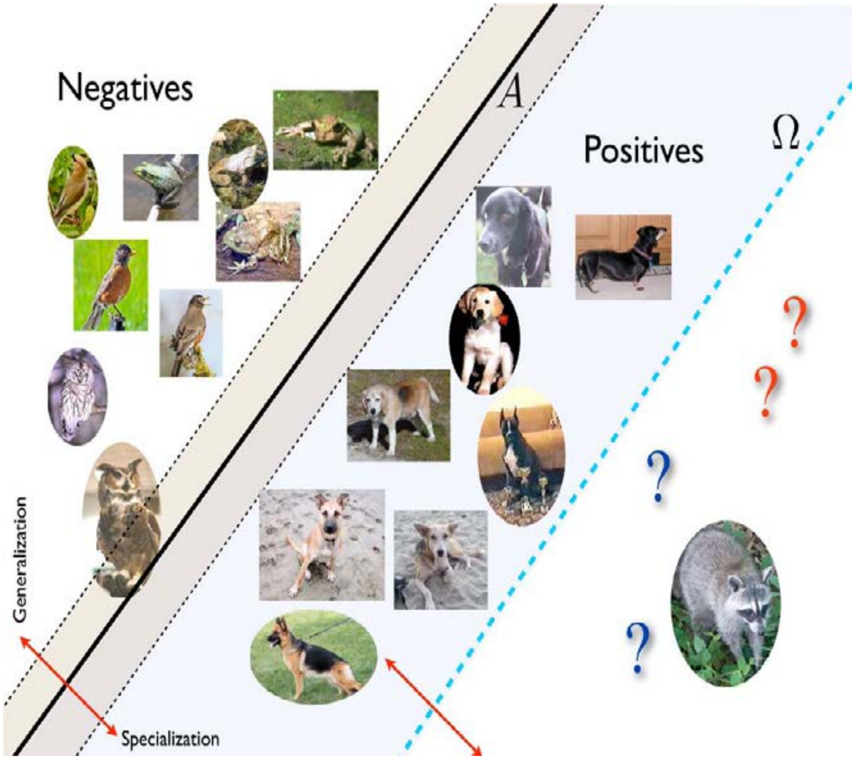
Biometric Challenges - 1

- Covariance Shift ~ Age-Pose, Illumination, and Expression (A-PIE) Variability
- Uncontrolled Settings and Media in the Wild
- Image Quality, Lack of Annotation, and Interoperability
- Alignment / Correspondence / Registration and Matching
- Occlusion, Disguise, Uncertainty ~ Graceful Degradation
- Spoofing and Liveness Detection
- Direct Learning and **Open Set Recognition**
- Reverse Learning and Contents-Based Image Retrieval (CBIR)

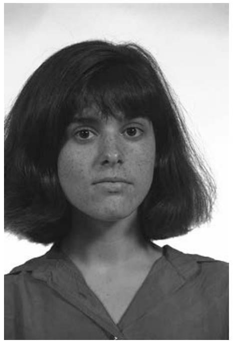
Biometric Challenges - 2

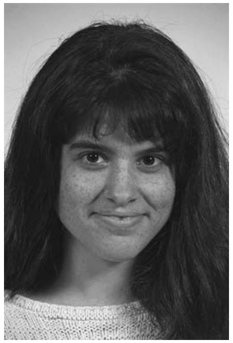
- Smart Identity Management ~ Anonymity, De-Duplication, and Privacy
- High-Dimensional Data and Scalability, Parameter Settings, and Score Normalization
- Biometric Data Sets / Demographics and Diversity / , Performance Metrics, Protocols, and Standards
- Performance Evaluation and Replication of Results
- Public Policy / Acceptance and Regulation, Social Contract, and Vulnerabilities
- Meta-Question: (a) How many different people are in set S?; (b) does subject A appear in S?

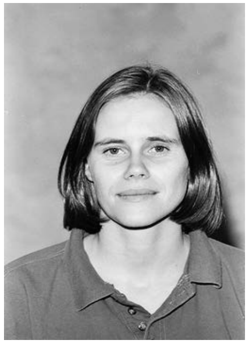
Open Set Recognition
- Scheirer et al., 2013 -

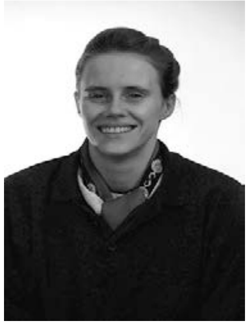


Intra-Class Variations





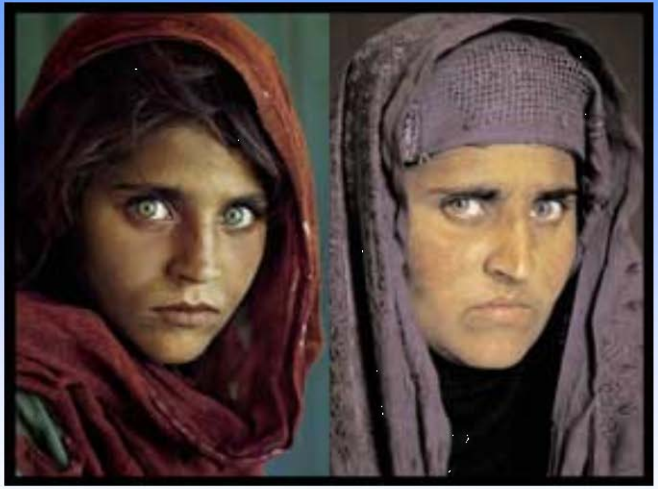




Inter - Class Similarity

Aging

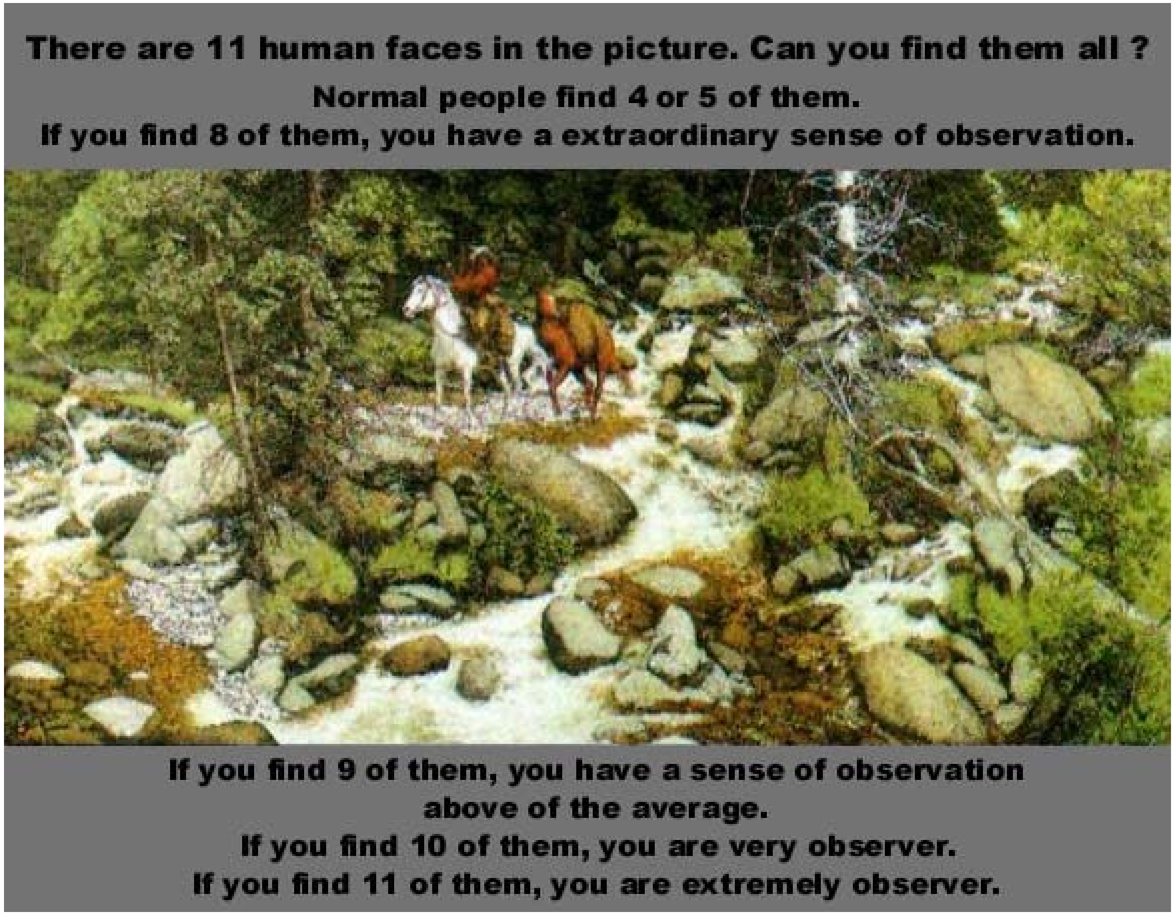
Sharbat Gula



1985 2002

(Left photo © Steve McCurry. Right photo Steve McCurry, © National Geographic Society.)

FACE IN A CROWD



There are 11 human faces in the picture. Can you find them all ?

Normal people find 4 or 5 of them.

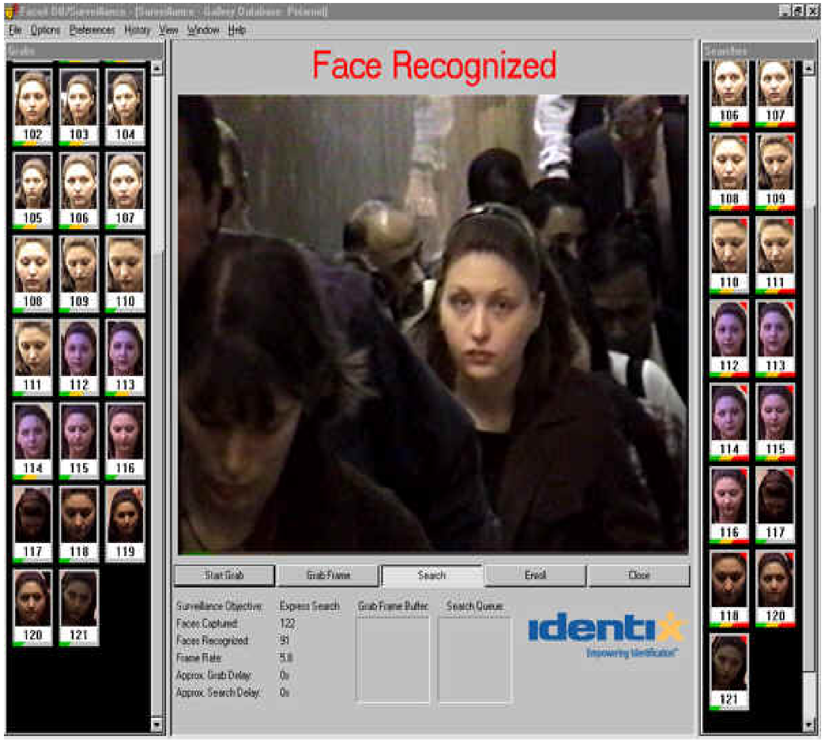
If you find 8 of them, you have a extraordinary sense of observation.

If you find 9 of them, you have a sense of observation above of the average.

If you find 10 of them, you are very observer.

If you find 11 of them, you are extremely observer.

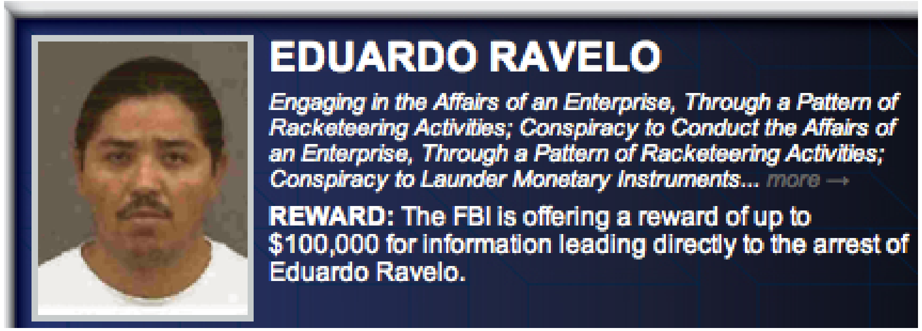
CCTV: link analysis, face selection, and re-identification



Surveillance Objective: Express Search
Faces Captured: 122
Faces Recognized: 91
Frame Rate: 5.8
Approx. Grab Delay: 0s
Approx. Search Delay: 0s

WANTED

FBI's Most Wanted Fugitive

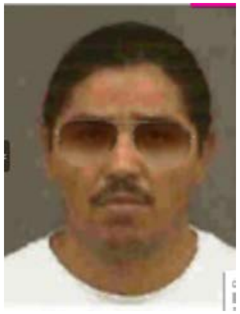


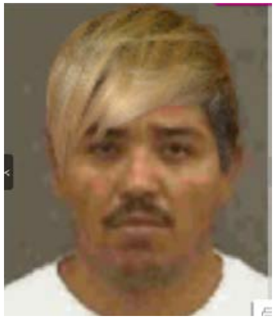
EDUARDO RAVELO

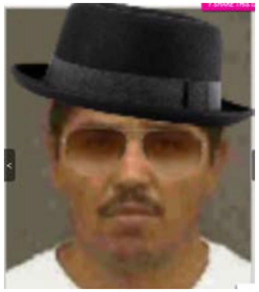
Engaging in the Affairs of an Enterprise, Through a Pattern of Racketeering Activities; Conspiracy to Conduct the Affairs of an Enterprise, Through a Pattern of Racketeering Activities; Conspiracy to Launder Monetary Instruments... more →

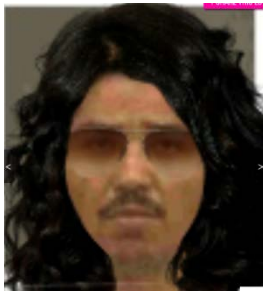
REWARD: The FBI is offering a reward of up to $100,000 for information leading directly to the arrest of Eduardo Ravelo.

Is this Eduardo?





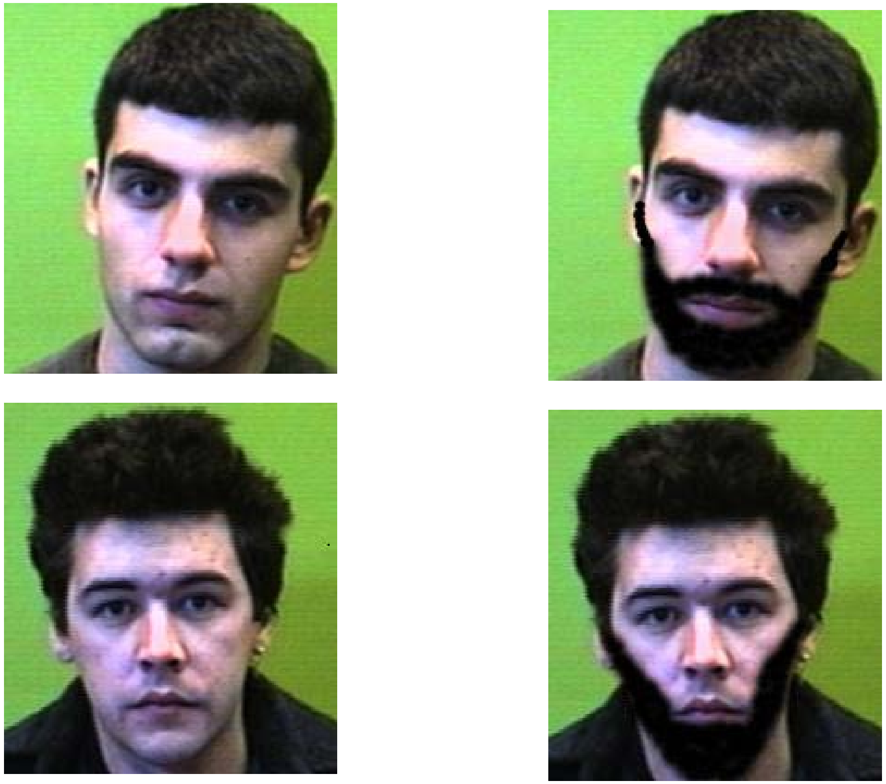




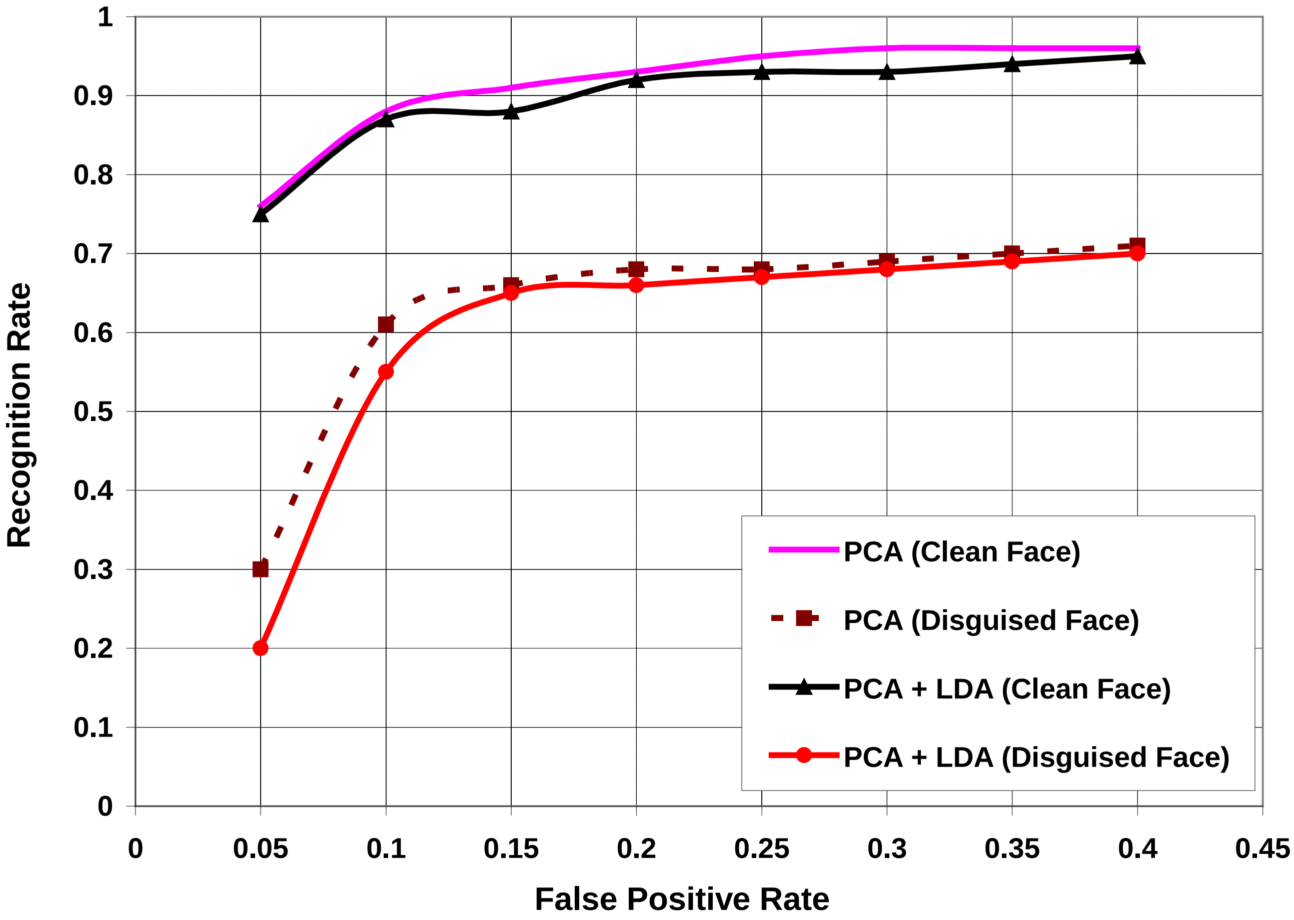
State of the art face recognition methodologies unable to identify imposters in the presence of occlusion and disguise.

Disguise

Sample Images (Clean Vs Disguised)



ROC for Disguise



Food for Thought

- **Data Driven --** In God we trust, all others must bring data (W. Edwards Deming) ☺
- **Competitive and Deep Learning** – Dictionaries , Sparse Coding (“**E**”), and Discrimination
- **Semantic-Spatial-Temporal Coherence --** (Multi-View) Evidence Accumulation, Consensus / Semi-Supervised Learning / Transduction / Recognition and Tracking / Multi-Label Collective & Iterative Classification / Conformal Prediction (“**M**”)
- **Recognition-by-Parts** – Region-Based Recognition / Qualitative Dipoles (Balas and Sinha), Data Fusion and Voting Methods (Balas and Sinha)
- **Transfer Learning** -- Learning with Side Information and from Auxiliary Tasks (“Helper Data”) – Multi-Task Learning

GESTALT - 1

- **Fusing the rich spatial, temporal, and contextual information available from the multiple views captured by today's "media in the wild."**
- Architecture ~ Configuration and Integration (of Labeled and mostly Unlabeled Data)
- "Helper" Data ~ Anchors and Transformations
- Distributed Memory ~ Snippets and Parts ~ LSH
- Evidence Accumulation and (loopy) Belief Propagation
- Multi-Layered Consensus / mid-layer vision functions {pose and segmentation}, Context (CRF, Huang et al., 2008) / and Semi-Supervised Learning

Sparse Coding

- Sparse representations for FR using L_1 minimization (Wright et al., 2009) presume "perfect registration, no self-shadowing, occlusion, or specularities."
- The sparsity assumption, which underpins much of FR is not supported by data (Shi et al., 2011), and does not improve recognition performance (Rigamonti et al., 2011). Sparsity and stability (important for generalization) are at odds with each other, and L_1 – regularized regression (Lasso) cannot be stable, and L_2-regularized regression is stable but not sparse (Xu et al., 2012).
- $\mathbf{A} = [x_1, \dots, x_n] \in \boldsymbol{R}^{m\,x\,n}$, $\boldsymbol{\Phi} \in \boldsymbol{R}^{d\,x\,m}$ (where $d \ll m$)
- $\boldsymbol{min}_{\alpha \in R^n} ||\boldsymbol{\alpha}||_1$, $\boldsymbol{\Phi x} = \boldsymbol{\Phi A \alpha}$ or equivalently $||\boldsymbol{\Phi x} - \boldsymbol{\Phi A \alpha}\,||_2 \leq \boldsymbol{\varepsilon}$
- $\boldsymbol{argmin}_{\alpha \in R^n} ||\mathbf{x} - \mathbf{A}\boldsymbol{\alpha}\,||_2^2$ with $\boldsymbol{\alpha} = (\mathbf{A}^{\mathrm{T}}\mathbf{A})^{-1}\mathbf{A}^T$
- $min_{D,\alpha(i)} \sum_i ||\mathrm{D}\alpha^{(\mathrm{i})} - \mathbf{x}^{(\mathbf{i})}||_2^2 + \lambda||\alpha^{(\mathrm{i})}||_1$ and $||D^{(j)}||_2^2 = 1$
- Supervised Sparse Coding // Restoration and Discrimination (Mairal et al., 2008)

Deep Learning

- Neocognitron (Fukushima) and Feed-Forward / (Local RF) Convolutional Networks // Distributed Representations // Model and Structure-- Hubel and Wiesel (V1 and V2, Olshausen and Field, LeCun, Bengio, Hinton, Ng --
- Hand-Crafted vs. Automatically (Unsupervised) Learned (Low- and High Level) Image Descriptors – (Back-Propagation and Loss Functions), Auto-Encoders, RBM, Deep Belief Networks (DBN)
- Dictionary (W), Encoder C (X, K), Filters (K), Sparse Coding (Z), Unsupervised Training, and Predictive Sparse Decomposition (PSD) --min $E(Z, W, K) = || X - WZ ||_2 + \lambda ||Z||_1 + ||Z - C(X, K)||_2$ (LeCun)
- Metric Learning – ITML, LMNN, CSML, MDS, ..
- Bottom-Up // Data Driven // Self-Taught (Unsupervised) Learning
- Grandmother Neurons ;-) Cat and Human (Face) Detectors

Model-Based Recognition

- **Face Space**

- What Is a Face?
- Anchors // global Dictionary and local kernels / and Transformations
- Training vs. Encoding

- **Learning from Data**

- Categorization and Generalization
- Data and Model Driven
- Evidence Accumulation and Belief Propagation
- Consensus
- Model Selection and Prediction

What Is a Face?

- HOG, LBP, Gabor, SIFT // (unsupervised) PCA, LDA, Fisherfaces, ICA //
- Neocognitron, Convolutional Networks, Auto-Encoding, Deep Learning
- The importance of (non-linear) Encoding (Sparse Coding) vs. (Unsupervised) Training (VQ and Random Exemplars)(Coates and Ng)
- Learning with Side (“Helper”) Information (“Similes”)(“Image Priors” /LDA Topics/ and Video Summarization) and Shepard’s 2nd Order Isomorphism (1968) → **Representation is Representation of Similarity** (Edelman, 1998) // Chorus of Prototypes
- Kernel Representation and Kernel Classifiers

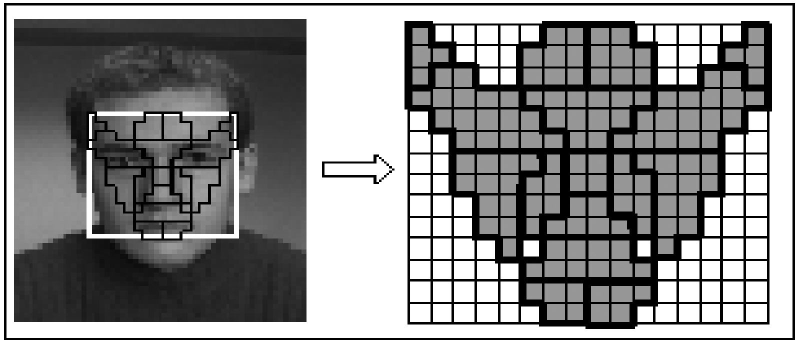
$$f(x) = \Sigma w(i) K_{\alpha}(x(i), x) \,/\, \Sigma\, K_{\alpha}(x(i), x) \;\&\; y = \Sigma_{i \in L} \alpha_i y_i K(x(i), x)$$

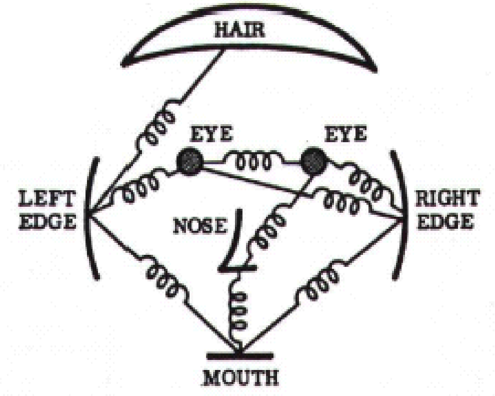
- Competitive Learning and Self-Organization ~ Vector Quantization (VQ), Code Book, and Semantic Networks (“Demographics”)

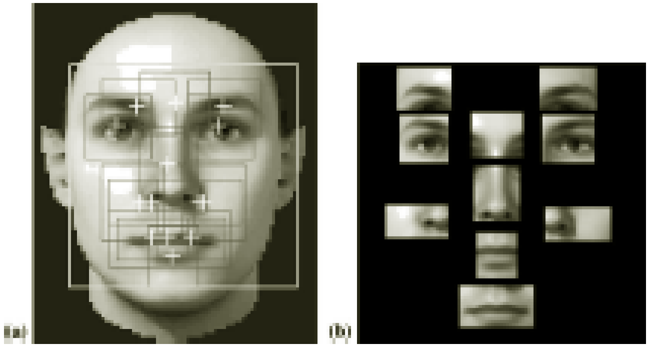
ART

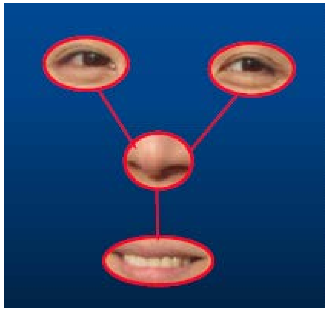
http://cs.gmu.edu/~wechsler/face-art.ppt

Recognition-by-Parts



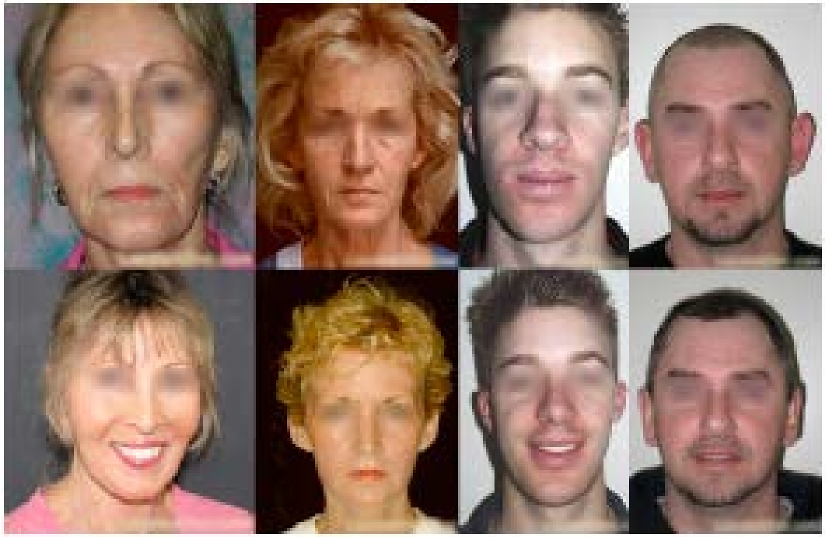




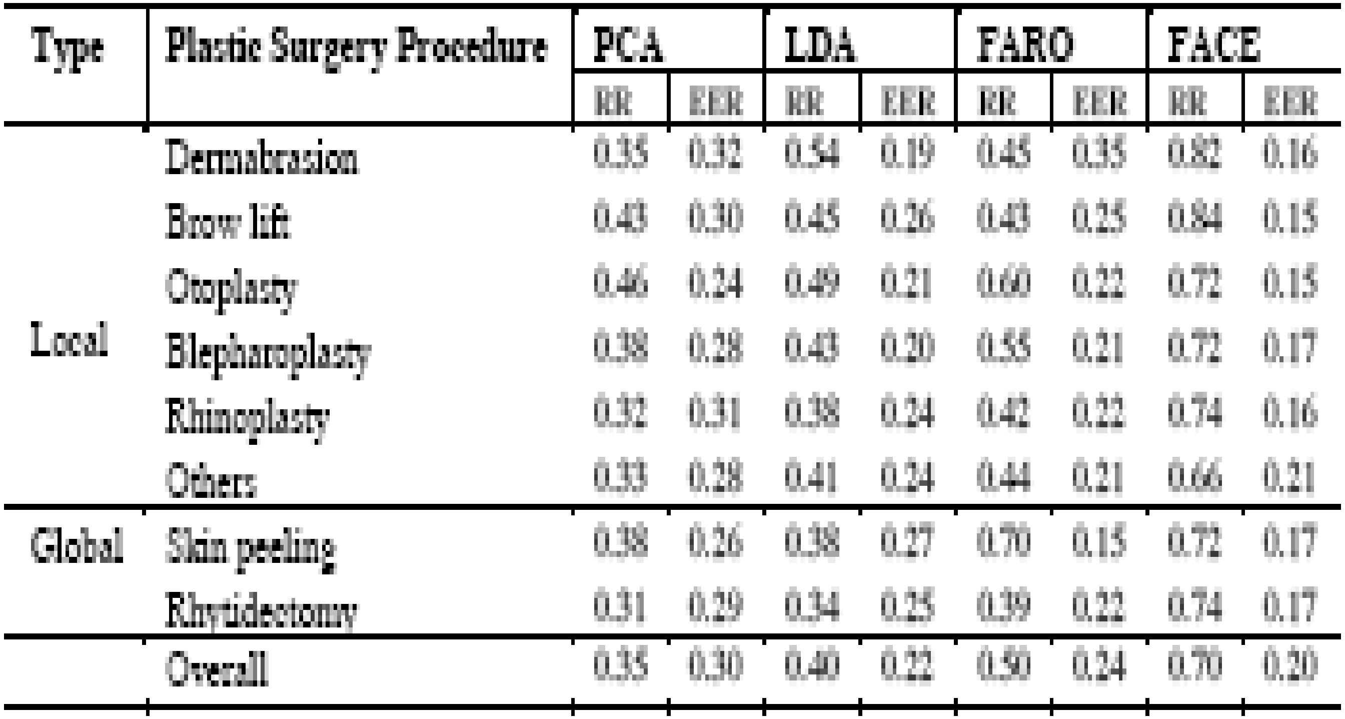


Face Recognition after Plastic Surgery

- DB (Singh et al.) : 1800 images // 900 subjects



Robust FR after Plastic Surgery Using Region-Based Methods - De Marsico, Nappi, Riccio, and Wechsler-



Type	Plastic Surgery Procedure	PCA		LDA		FARO		FACE	
		RR	EER	RR	EER	RR	EER	RR	EER
Local	Dermabrasion	0.35	0.32	0.54	0.19	0.45	0.35	0.82	0.16
	Brow lift	0.43	0.30	0.45	0.26	0.43	0.25	0.84	0.15
	Otoplasty	0.46	0.24	0.49	0.21	0.60	0.22	0.72	0.15
	Blepharoplasty	0.38	0.28	0.43	0.20	0.55	0.21	0.72	0.17
	Rhinoplasty	0.32	0.31	0.38	0.24	0.42	0.22	0.74	0.16
	Others	0.33	0.28	0.41	0.24	0.44	0.21	0.66	0.21
Global	Skin peeling	0.38	0.26	0.38	0.27	0.70	0.15	0.72	0.17
	Rhytidectomy	0.31	0.29	0.34	0.25	0.39	0.22	0.74	0.17
	Overall	0.35	0.30	0.40	0.22	0.50	0.24	0.70	0.20

ARCF (Hung, Ramanathan, and Wechsler)
- adaptive and robust correlation filters -

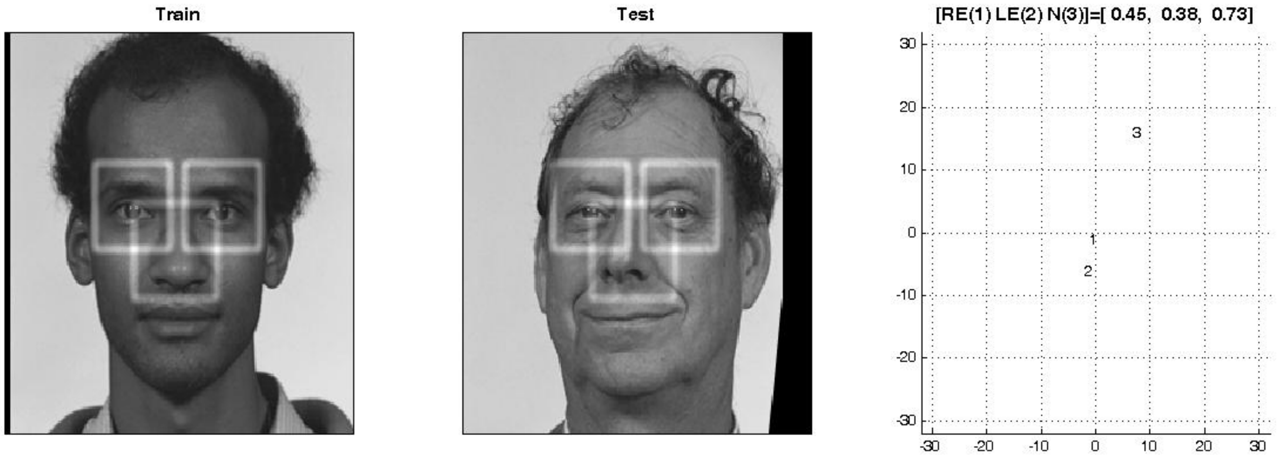
Occlusion/disguise not necessarily deliberate –aging, hair style, injuries,..

Recognition-by-parts

Uses both training and test data ~ Adaptive and Robust Correlation Features (ARCF)

Less sensitive to noise and distortions

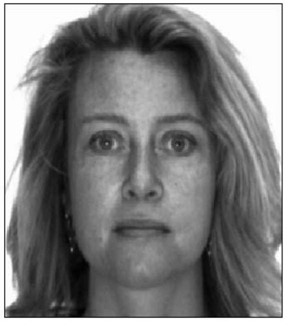
Adaptive and Robust Correlation Filters -
- Ventral ~ *what* ~ and Dorsal ~ *where* ~ Paths -



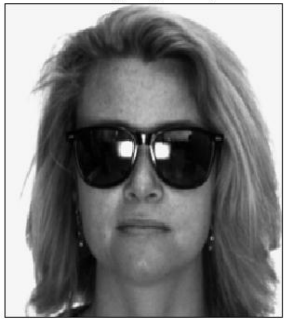
Occlusion

Shift-Invariance, Graceful Degradation, and Closed-Form Solutions

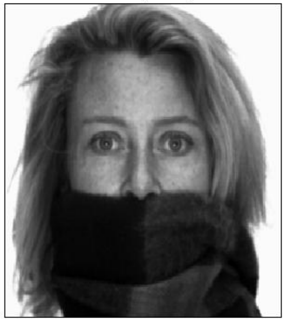
Train



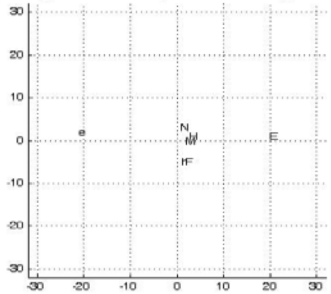
Test1 (Sunglasses)



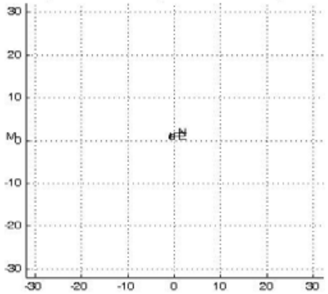
Test2 (Scarf)



[F H P]-thresholds=[-0.02 0.02 0.13] Accept



[F H P]-thresholds=[0.11 0.09 0.16] Accept



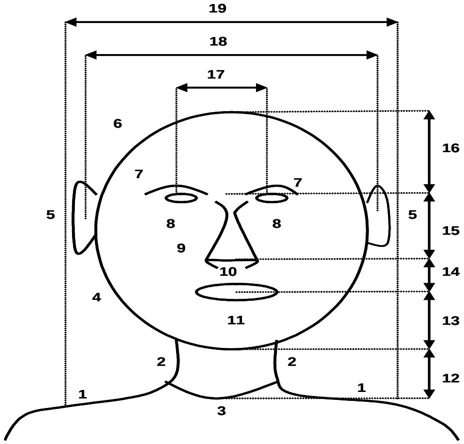
Hybrid (Anthropometric + Appearance) Representation (Ramanathan and Wechsler)

Novelties - hybrid feature sets and fusion methodology

Anthropometric : Head, face and shoulder, linear and non-linear measurements

Appearance : PCA/PCA+LDA Eigen vectors

Feature level fusion and decision level fusion for identification



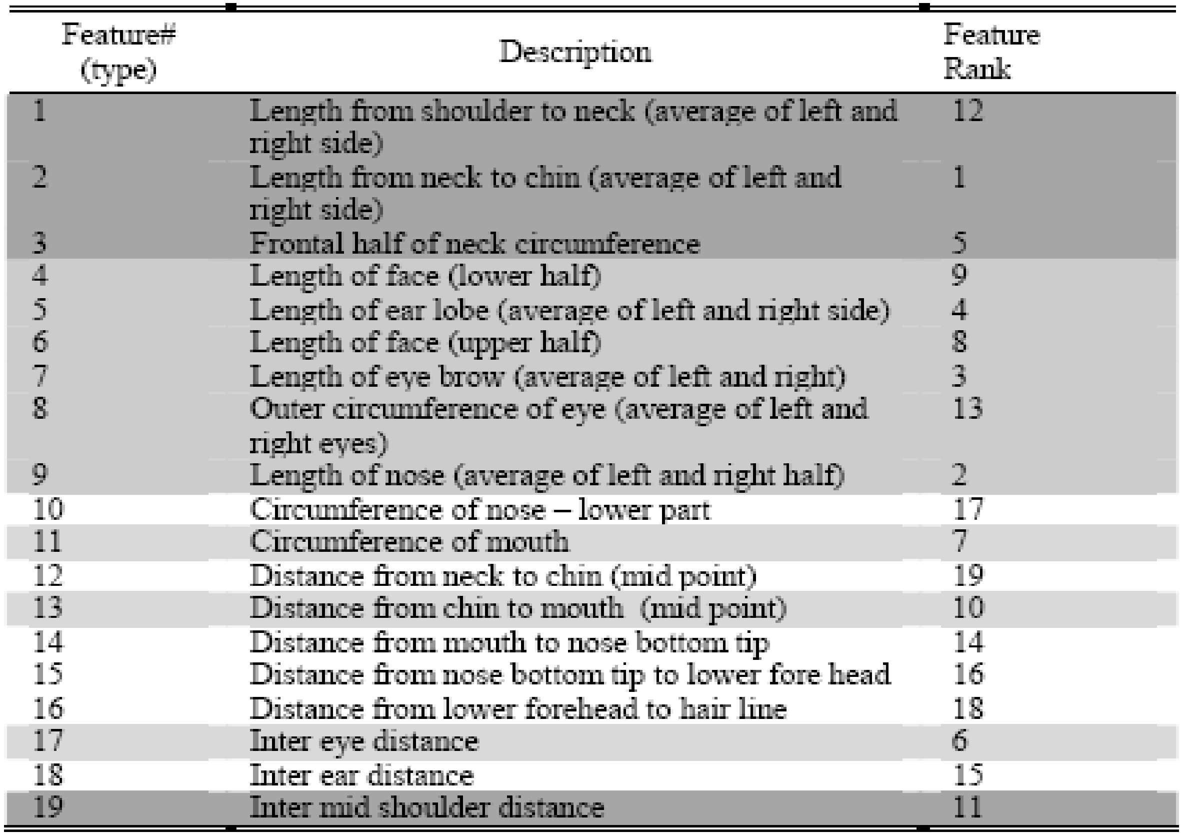
Anthropometric measurements

+

PCA/PCA+LDA Eigen/Fisher faces

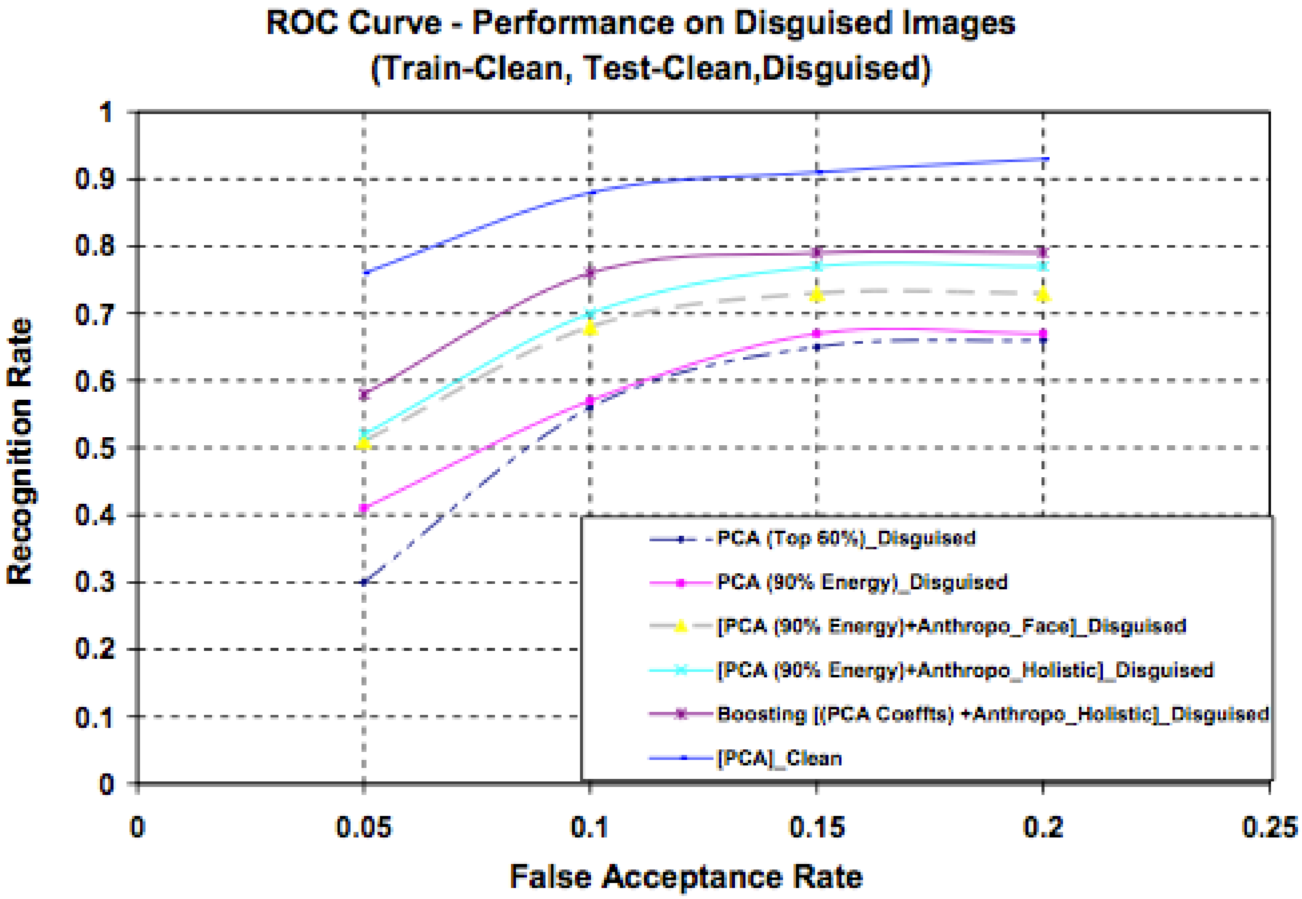
Mutual Information and Filter Approach

Table 1. Ranking of Anthropometric Features



Feature# (type)	Description	Feature Rank
1	Length from shoulder to neck (average of left and right side)	12
2	Length from neck to chin (average of left and right side)	1
3	Frontal half of neck circumference	5
4	Length of face (lower half)	9
5	Length of ear lobe (average of left and right side)	4
6	Length of face (upper half)	8
7	Length of eye brow (average of left and right)	3
8	Outer circumference of eye (average of left and right eyes)	13
9	Length of nose (average of left and right half)	2
10	Circumference of nose – lower part	17
11	Circumference of mouth	7
12	Distance from neck to chin (mid point)	19
13	Distance from chin to mouth (mid point)	10
14	Distance from mouth to nose bottom tip	14
15	Distance from nose bottom tip to lower fore head	16
16	Distance from lower forehead to hair line	18
17	Inter eye distance	6
18	Inter ear distance	15
19	Inter mid shoulder distance	11

Hybrid (Anthropometric + Appearance) Representation



CONSENSUS

- RANdom Sample Consensus (RANSAC), Hough Transform (HT), and Random Hough Transform (RHT)

- Semi-Supervised Learning (SSL) ~ Regularization (Local and Global Consistency (LLGC) / Information Diffusion / Random Walks (Zhu et al., 2003; Zhou et al., 2005) / Label Propagation / Augmented Graphs and External Classifiers /

- Model Selection / MDL ~ Transduction, strangeness / typicality **α** ~ ranking **p-values**

- Voting Methods ~ Bagging (“Random Forests”) and Boosting

- Reverse Learning ~ Indexing ~ CBIR ~ sub-linear time and Local Sensitivity Hash (LSH) (storage)

Semi – Supervised Learning (SSL) and Transduction

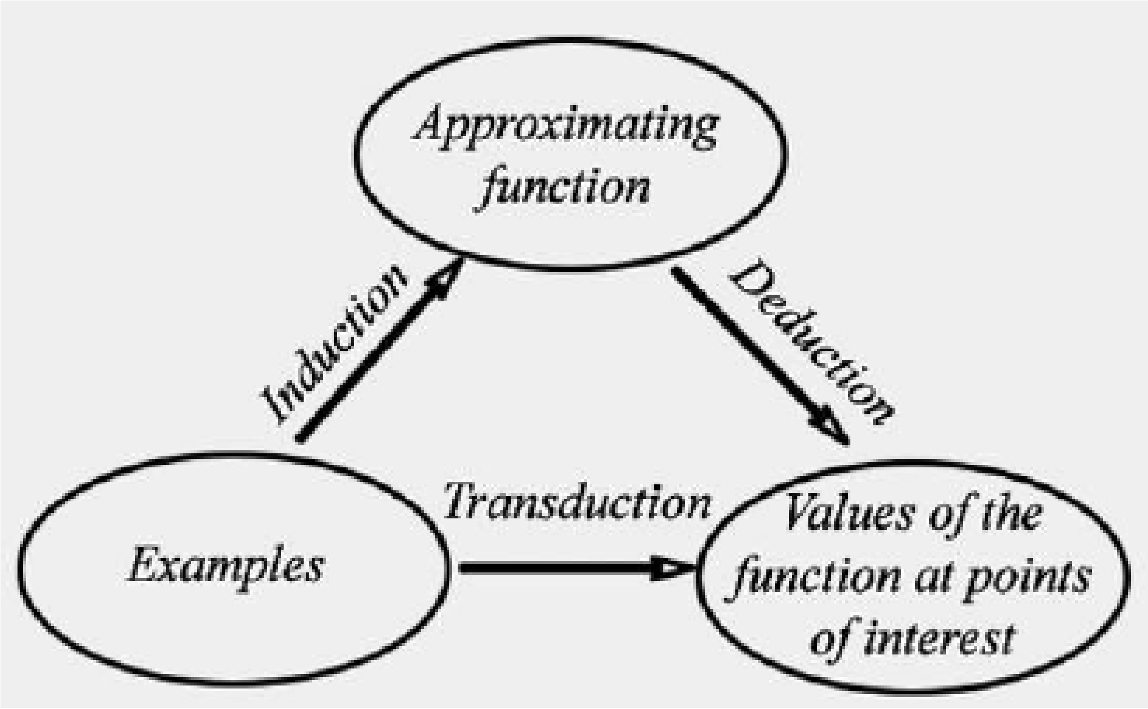
- **Transduction ~ maximize the margin over both labeled and unlabeled exemplars ~ training and test data are complementary ~ learning from unlabeled *ghost / virtual* samples ~ stability and consistency** ***~ take home exam ~ less human effort and better accuracy (Seeger) ~ unlabeled data provides <u>information</u> about the structure of the domain / underlying pdf, whereas the labeled data identifies the classification task / <u>robustness</u> within this structure.***

- **SSL ~ contradictions, hints, and inductive bias ("metrics")** ***~ access to previous exams ~ consistency and cluster assumption / boundaries lie in regions of low data density ~ manifold assumption / ?? distant data points are very unlikely to take similar labels ?? /***

- **Co-Training, Multi – Task Learning, and Transfer Learning ~** ***representation ("code") sharing***

Transduction

- strangeness and typicality -



Motivation for Transduction

- Labels for **T**raining and **t**esting are made compatible (vs. local / global // LLGC)
- **T**raining and **t**esting errors are consistent
- Learn data distribution from **t**est data
- Relative similarity scores and **rankings**
- **Adversarial Learning** – label flipping – **Consensus** – Outlier / **Imposter Detection** – **Open Set Recognition**

Conformal Prediction - 1

- Algorithmic Learning in a Random World (Vovk, Gammerman and Shafer, 2005).
- Conformal Prediction (CP) complements the predictions made by ML algorithms with metrics of reliability, e.g., non-conformity measures (NCM).
- The purpose for NCM is to support hedging / punting between accuracy and confidence, when making predictions, according to the costs and risks involved. In particular, the methods developed using the CP framework produce “well calibrated” reliability measures for individual examples without assuming anything more than that the data are generated independently by the same (but unknown) probability distribution (i.i.d).
- Transduction leverages non-confidence measures (NCM), makes use of both labeled (annotated) and unlabeled biometric data, addresses multi-layer categorization, and provide NCM of reliability in the predictions made, e.g., credibility and confidence. Transduction Confidence Machine for Detection and Recognition (TCM-DR) expands on the traditional Transduction Confidence Machine (TCM).

Conformal Prediction - 2

- The credibility measure is well-calibrated (or conservatively valid) as the frequency of prediction error does not exceed significance level ϵ (between 0 and 1) at a chosen confidence level $1 - \epsilon$ (in the long run). Smaller values of ϵ correspond to greater reliability. The confidence measure, which expresses the extent of ambiguity, becomes efficient as the TCM and TCM-DR prediction sets (regions) shrink (in terms of number of possible outcomes).

- The basic mode of operation for transduction is incremental in nature as it leverages the complementarity between training and test data for the purpose of robust and stable predictions. TCM and TCM-DR enable meta-prediction for learning from both labeled (annotated) and unlabeled examples, while employing ranking and sensitivity analysis for the purpose of sequential importance sampling (SIS), adversarial learning, and perturbations / revisions during on-line recognition and tracking.

Conformal Prediction - 3

- A predictor is said to be *valid* (or well calibrated) if its frequency of prediction errors does not exceed ϵ at a chosen confidence level $1 - \epsilon$ in the long run. In addition, a predictor is *efficient*, i.e., performs well, if the prediction set is as small as possible. A conformal predictor maps a data sequence *S*, a new object x, and a confidence level $1 - \epsilon \in (0, 1)$ to the prediction set

$$\Gamma^{\epsilon} (S, x) = \{y \in Y: p_y > \epsilon\} \text{ with } \Gamma^{\epsilon 1} \leq \Gamma^{\epsilon 2} \text{ for } \epsilon 1 \geq \epsilon 2$$

- The efficiency of the nested family of predictions Γ^{ϵ} is directly related to diffusion during tracking (using particle filtering). An empty prediction is a warning that the object to be predicted is unusual (the credibility is ϵ or less). The transduction conformal predictor is valid in the sense that the probability of error that a correct label y^c does not belong to $\Gamma^{\epsilon}(S, x)$ at confidence level $1 - \epsilon$ never exceeds ϵ.

- Sensitivity Analysis ~ Revision and Stability ~ Conformal Inductive Predictor (ICP) (Nappi and Wechsler)

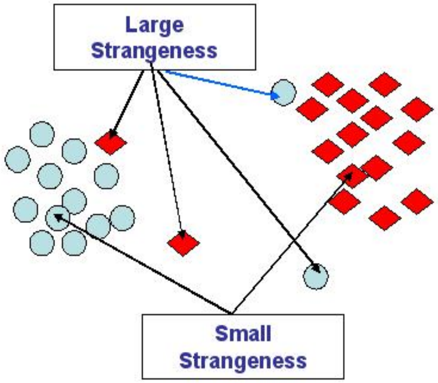
Non-Conformity Measures (NCM)

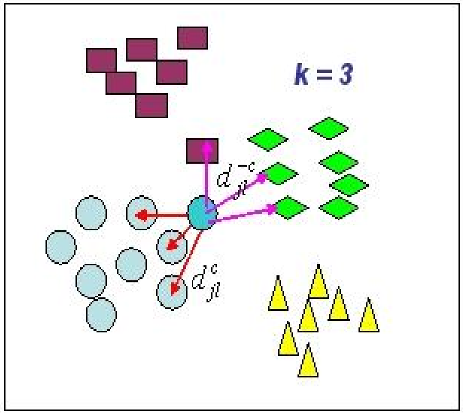
- Strangeness (Typicality)
- Hypothesis Margin ~

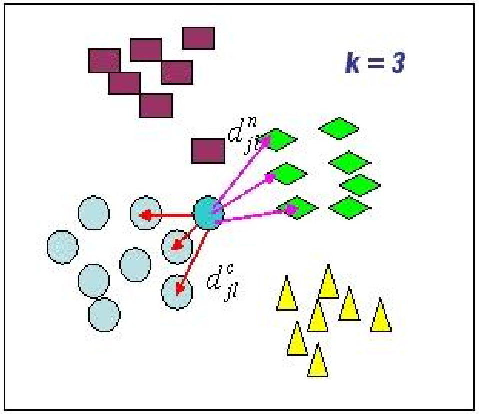
  $$\Phi(x) = (||x - \text{near-miss}(x)|| - ||x - \text{near-hit}(x)||)$$

- Randomness Deficiency
- p-values and Rankings

Strangeness







Strangeness [“Cohort”] Definitions

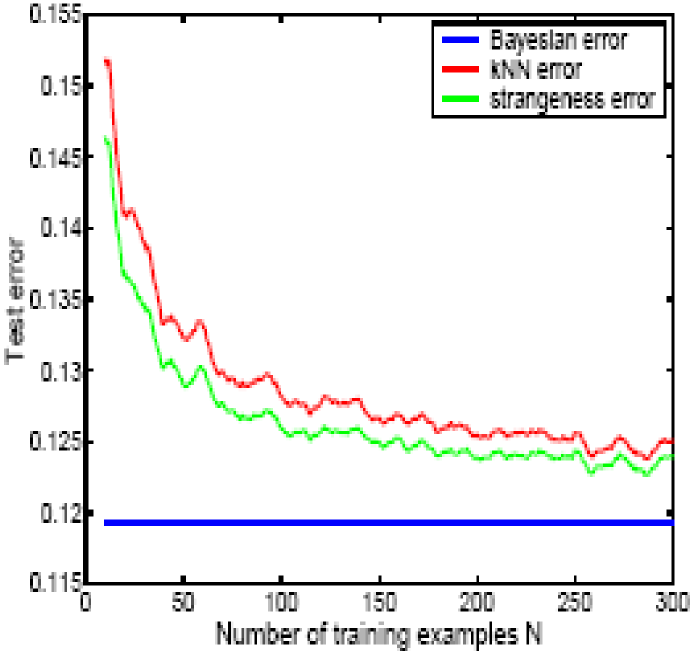
$$\alpha_i = \frac{\sum_{j=1}^{k} d_{ij}^{y}}{\sum_{j=1}^{k} d_{ij}^{\neg y}}$$

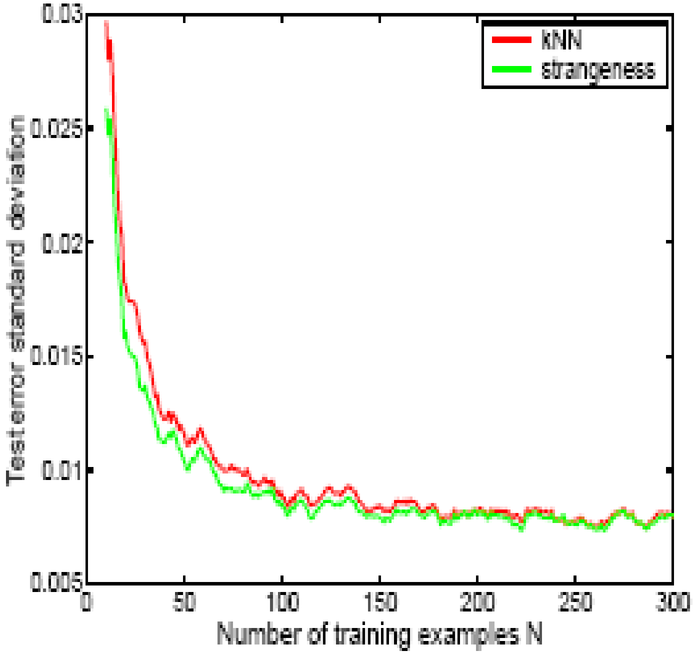
$$\alpha_j = \frac{\sum_{l=1}^{k} d_{jl}^{c}}{\min_{C, n \neq c} \sum_{l=1}^{k} d_{jl}^{n}}$$

Strangeness and Decision Boundaries

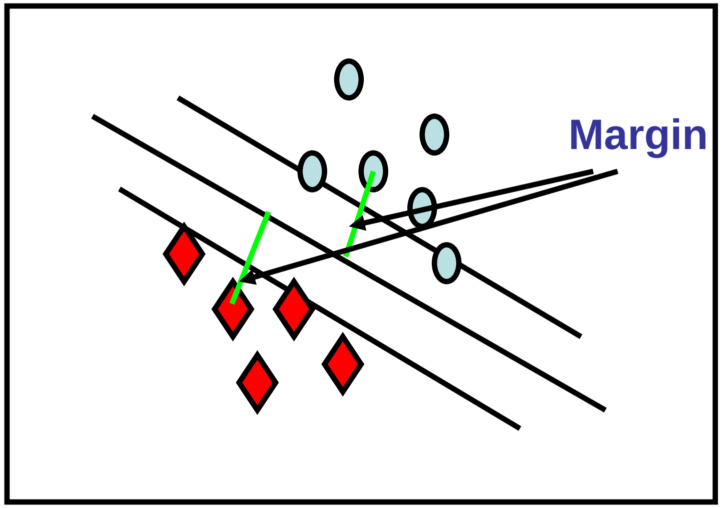
- ***k-NN*** error approaches the Bayes error (with factor 1) if ***k*** = *O* (log *n*)
- strangeness ***α*** is related to the optimal decision boundary (***α*** = 1) and the posterior P (c_j | x_i)
- ***k - NN* strangeness** smoothes boundaries and generalizes better than ***k - NN*** particularly for overlapping distributions

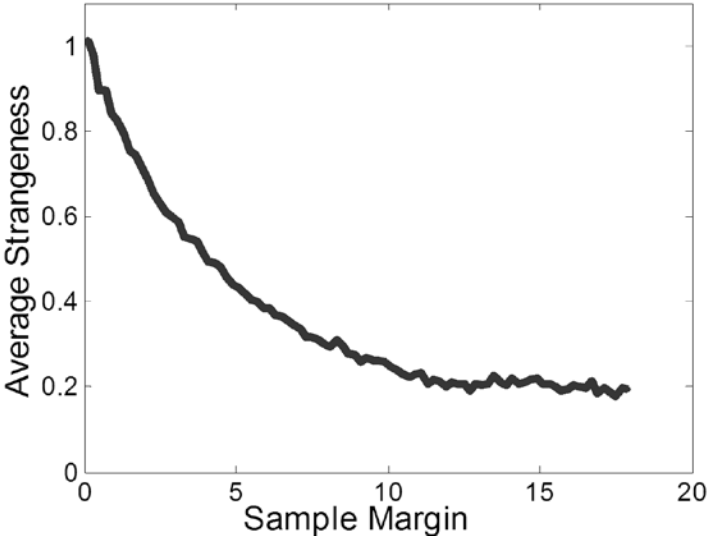
k-NN vs. k-NN strangeness





Margin and Strangeness





Kolmogorov Complexity and Randomness Deficiency

- Let *S* be the set of binary strings ***x*** of fixed length and Kolmogorov complexity K(**x**). The randomness deficiency *D (**x**|S)* for string ***x*** is

$$D\ (\mathbf{x}|S) = \log |S| - K(\mathbf{x}|S)$$

with D(**x**|S) a measure of how random the binary string ***x*** is. The larger the randomness deficiency is the more regular and more probable the string ***x*** is.

- Kolmogorov complexity and randomness are conceptually related through the minimum description length (MDL).

Randomness Deficiency
- computation -

- Randomness deficiency is not computable (Li and Vitanyi, 1997).
- Martin – Löf test for randomness using ***p – values*** (different from those used in statistics to support or discredit the null hypothesis)
- ***p – values*** are defined using the strangeness
- large random deficiency ~ large ***p – values*** ~ *typical examples*

p – values - for putative label *y* -Transduction Confidence Machine (TCM) and Open Set Recognition

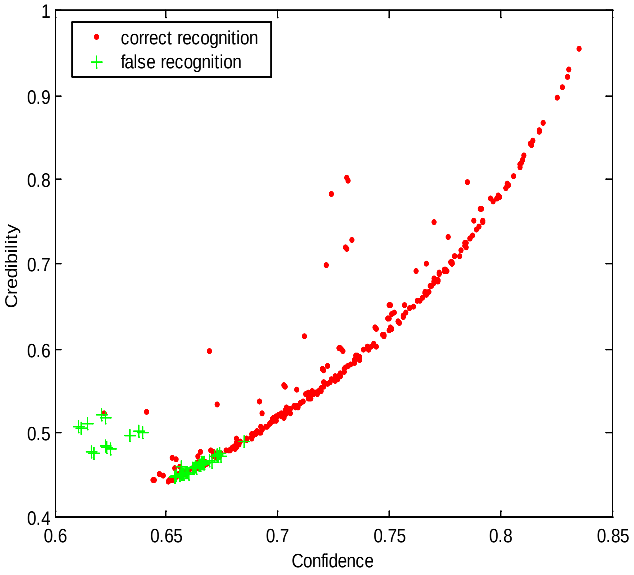
$$p_y(e) = \frac{\#\{i : \alpha_i \geq \alpha_{new}^y)}{l+1}$$

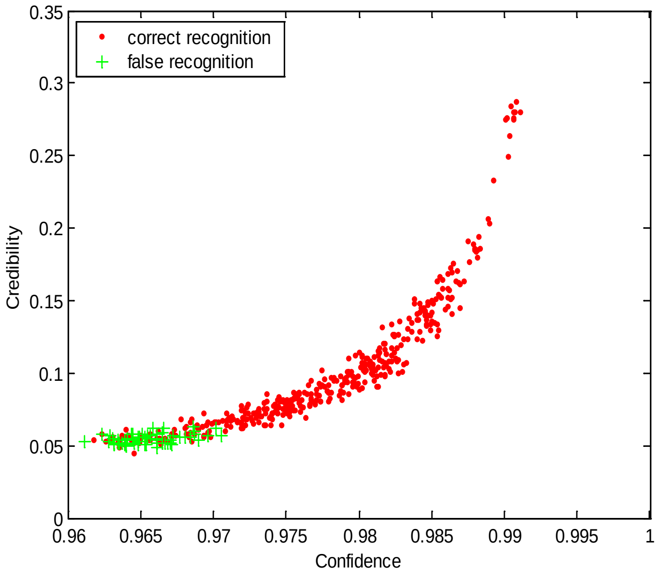
$$p_y(e) = \frac{f(\alpha_1) + f(\alpha_2) + \cdots + f(\alpha_l) + f(\alpha_{new}^y)}{(l+1)\, f(\alpha_{new}^y)}$$

p-values
- credibility and confidence -

p-values determine the relative strangeness (or typicality). [extensive testing shows that the 2nd strangeness definition provides better performance.] The largest p-value defines the credibility of the classification chosen. The confidence measure is the difference between the top two p-values. It indicates how close to each other the first two classifications and it measures ambiguity. Credibility and confidence are examples of information quality. The distribution of p-values defines PSR values and is used to detect and reject outliers.

PCA and Fisherfaces [FERET]
- credibility and confidence -





Open Set Recognition Using Transduction

1. Strangeness (“typicality”) computation

2. Randomness deficiency and p-values derivation

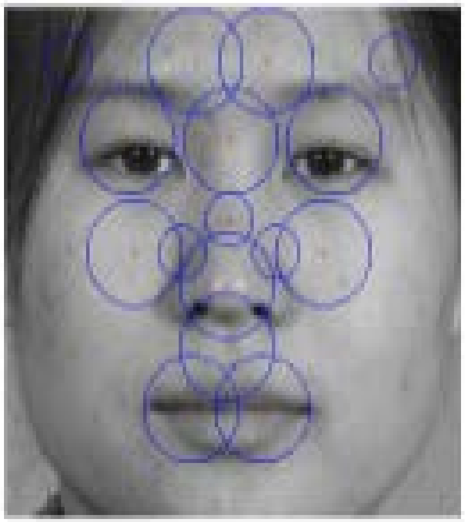
3. A priori setting for [impostor] thresholds and detection decisions (for outlier rejection) in terms of [low] p-values using PSR (peak-side-ratio [PSR]) = $(p_{max} – p_{mean}) / p_{sd}$

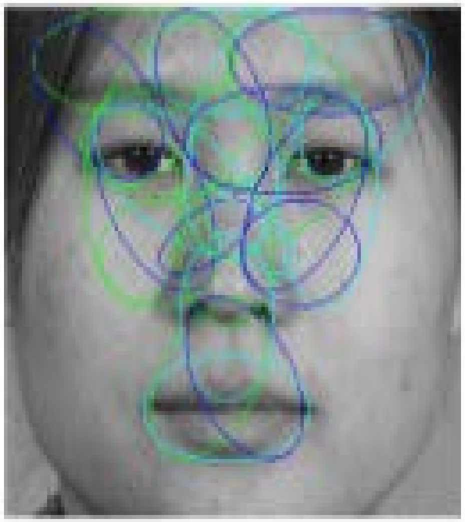
4. Open Set TCM - kNN (Transduction Confidence Machine – k Nearest Neighbors) recognition algorithm

- Representation and Classification –

- Scale Invariant Feature Transform (SIFT) and Patch Representation
- Feature (and Variable / Dimensionality) (*Patch*) Selection Using Strangeness and Iterative Backward Elimination [mutual information and Markov blankets]
- From Patches to (exemplar based) Parts Using K – Means Clustering
- *Weak* Learners ("Parts") Compete to Assembly *Strong* Learners by Boosting

1st and 2nd Order Patches

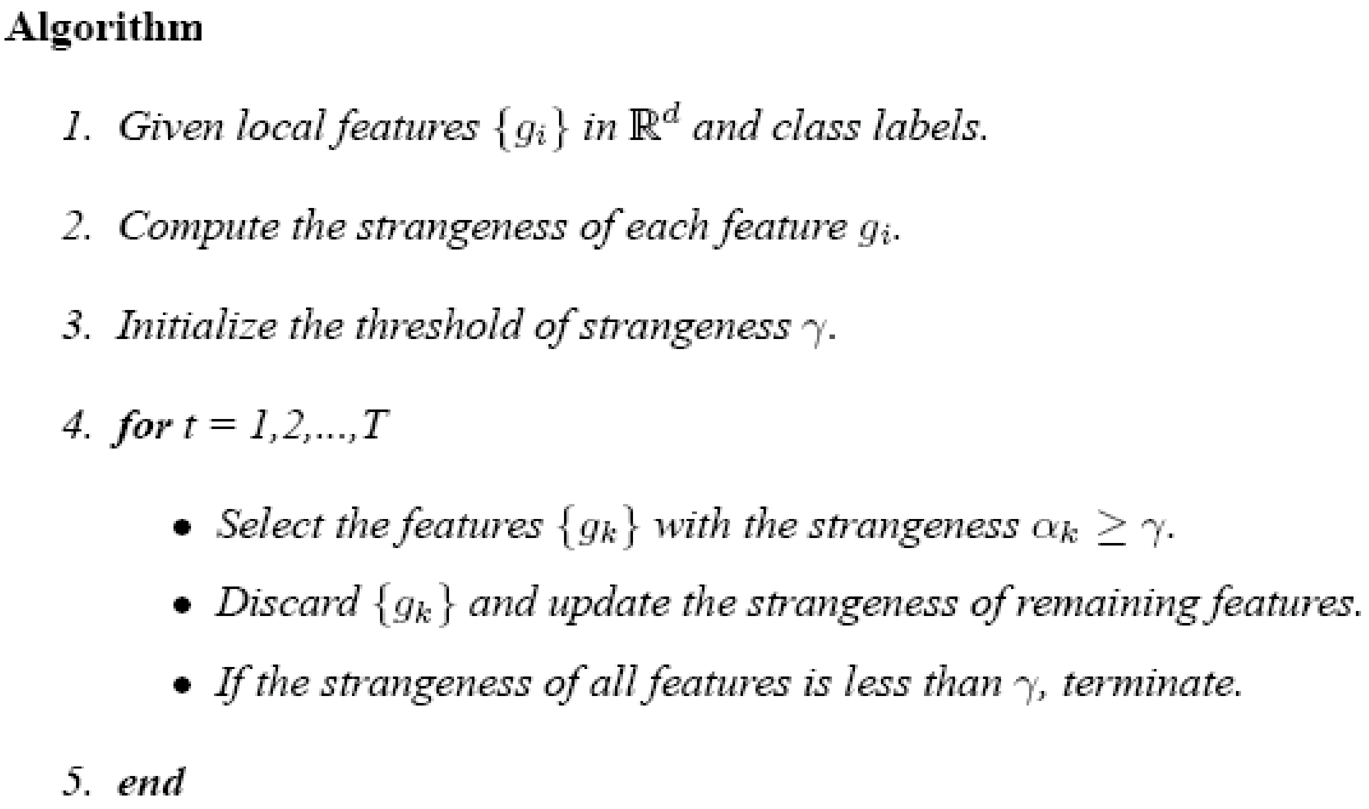




Feature (“x”) Selection and Dimensionality (“y”) Reduction - feature ~ patch -

- Iterative Backward Elimination ~ Markov Blanket Filtering ~
- remove irrelevant features <RELIEF>, remove redundancies (K-Means), and combinatorial feature selection.

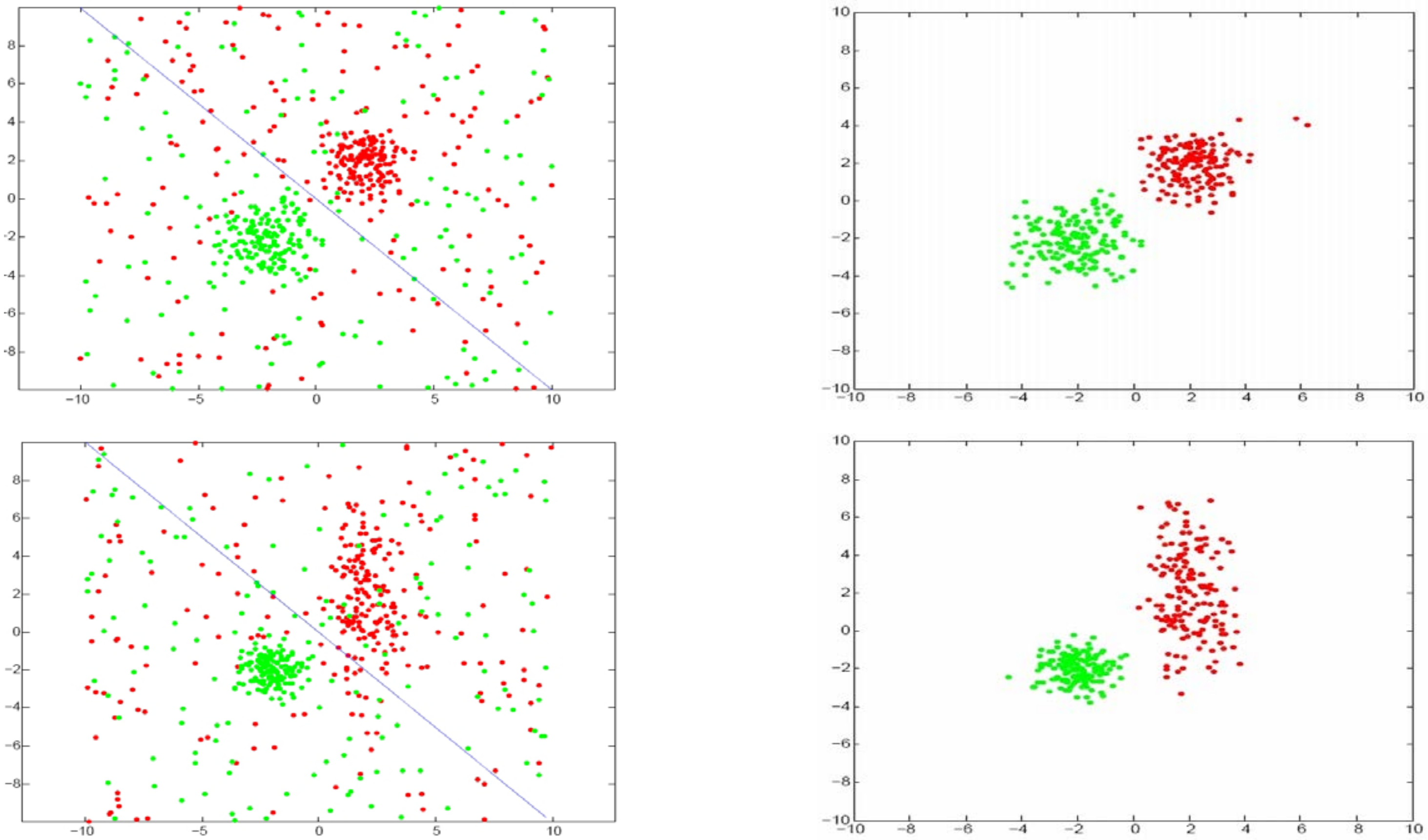
Feature Selection Using Strangeness



Algorithm

```
1. Given local features $\{g_i\}$ in $\mathbb{R}^d$ and class labels.
2. Compute the strangeness of each feature $g_i$.
3. Initialize the threshold of strangeness $\gamma$.
4. for $t = 1,2,\ldots,T$
      • Select the features $\{g_k\}$ with the strangeness $\alpha_k \geq \gamma$.
      • Discard $\{g_k\}$ and update the strangeness of remaining features.
      • If the strangeness of all features is less than $\gamma$, terminate.
5. end
```

Feature Selection Results

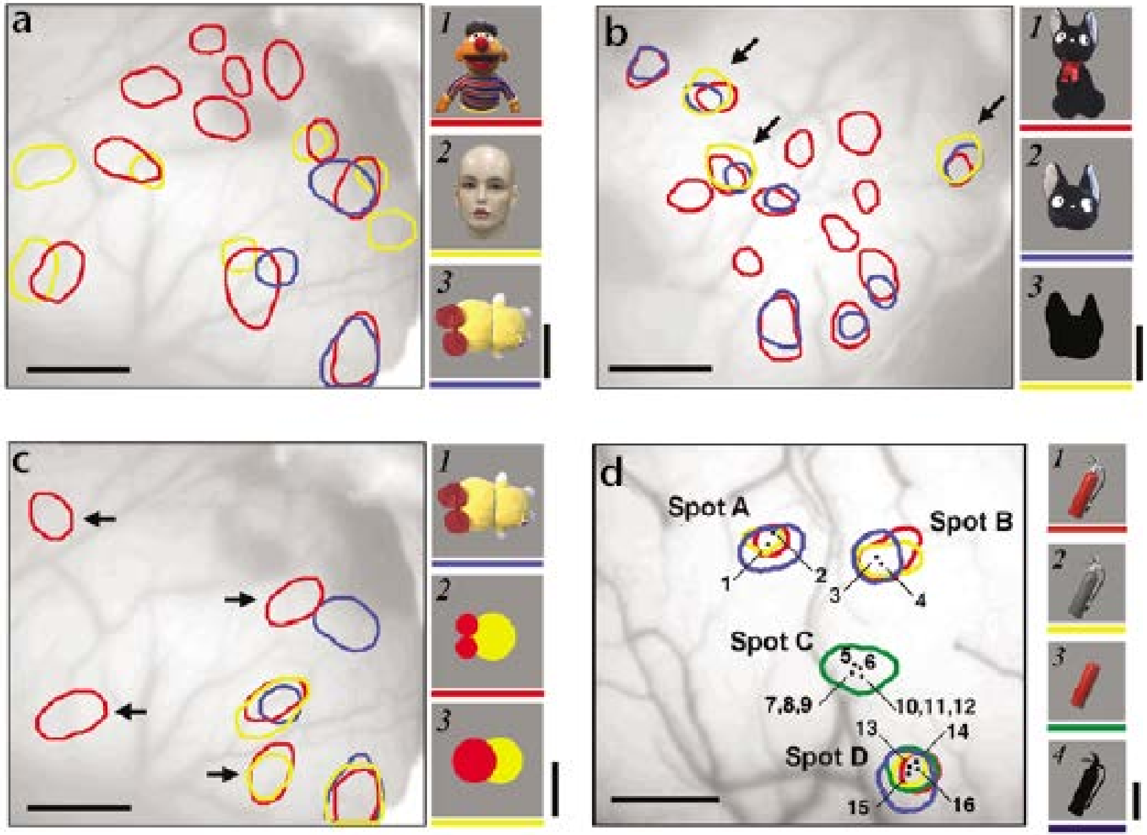


Cortex Representations
- distributed representations -

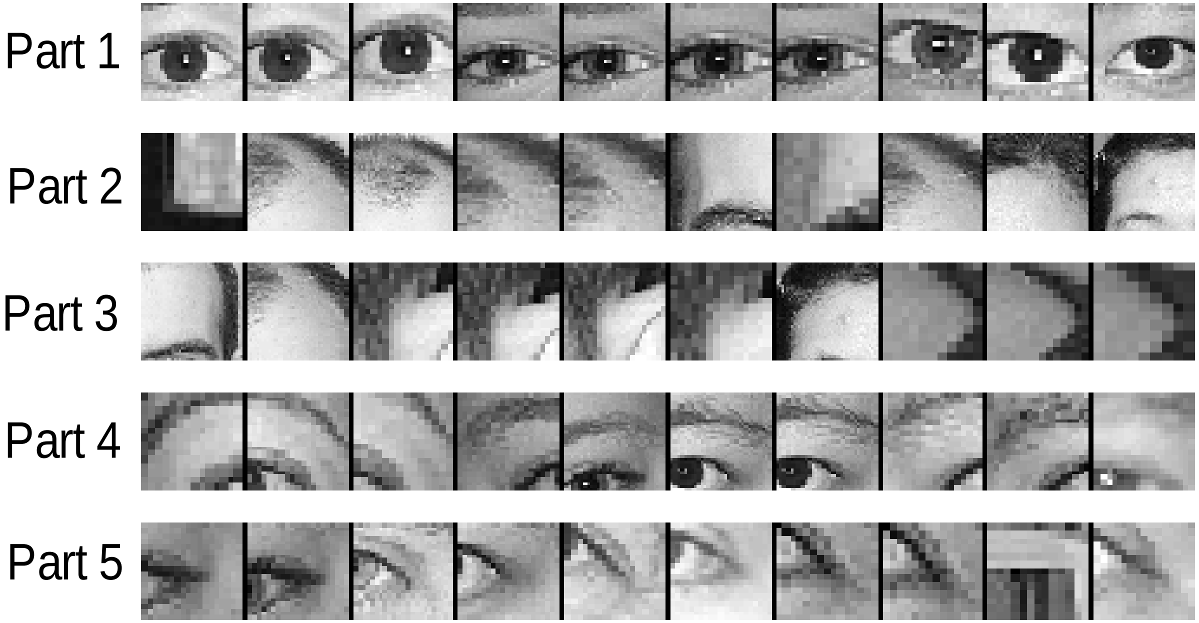
Complex objects are represented in macaque IT cortex by the combination of feature columns...These results suggest that objects may be represented not only by simply combining feature columns but also by using a variety of combinations of active and inactive columns for individual features (Tsunoda et al., 2001) [sparse codes for association ~ Barlow, 1989]

Over-Complete dictionaries, [Hausdorff distance], flexible matching and redundancy reduction

(Tsunoda et al., 2001)
V1 → IT → TE



From Patches (“Features”) to Parts Using K- Means Clustering



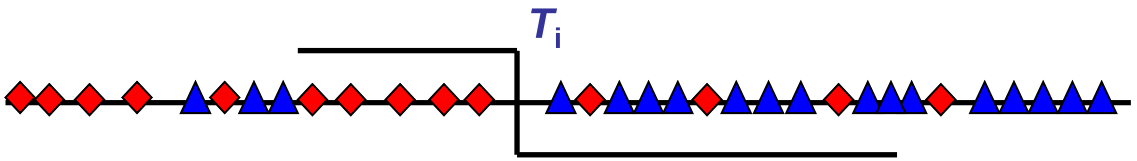
Boosting and Validation

- Recognition by Parts Using Boosting and Transduction – Li and Wechsler-

- Build a weak classifier for each part.
- Relative strangeness of the parts determines the threshold T_i and coefficients β_i of each weak classifier using validation.
- Choose best “part” and iterate
- AdaBoost.M2 ~ hard labels and hard examples

Strangeness Weak Learner

- C classes are given, each of which has N validation images.
- For each part of class c ($c=1, .. ,C$), there are N positive examples of strangeness values – one from each validation image of class c.
- For each part of class c, there are $N(C-1)$ negative examples of strangeness values – one from each validation image of other classes.
- Rank all the strangeness values and select the best threshold T_i to get the maximal recognition rate for those validation images

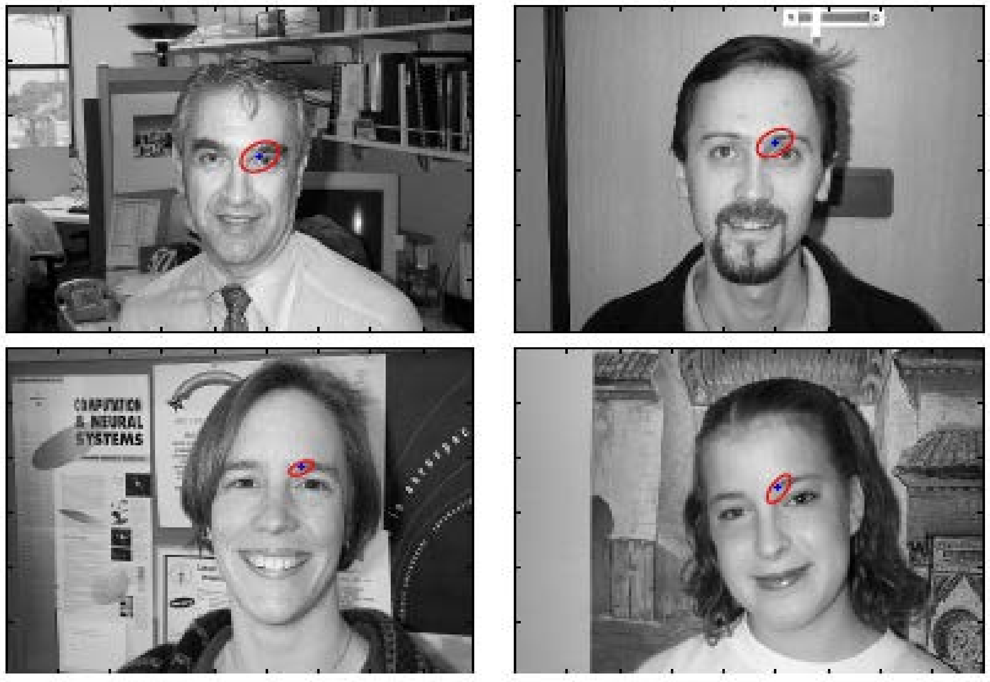


Strangeness values

◆ Strangeness values from validation images of class c

▲ Strangeness values from validation images of all other classes

Eyebrows – Best Part / Weak Learner for Categorization / “Detection” / Using Boosting



UND ~ FRGC Experiment

200 Subjects x 48 Images

12 (training) + 12 (validation) + 24 (testing)

Recognition Performance
Using Boosting and Transduction

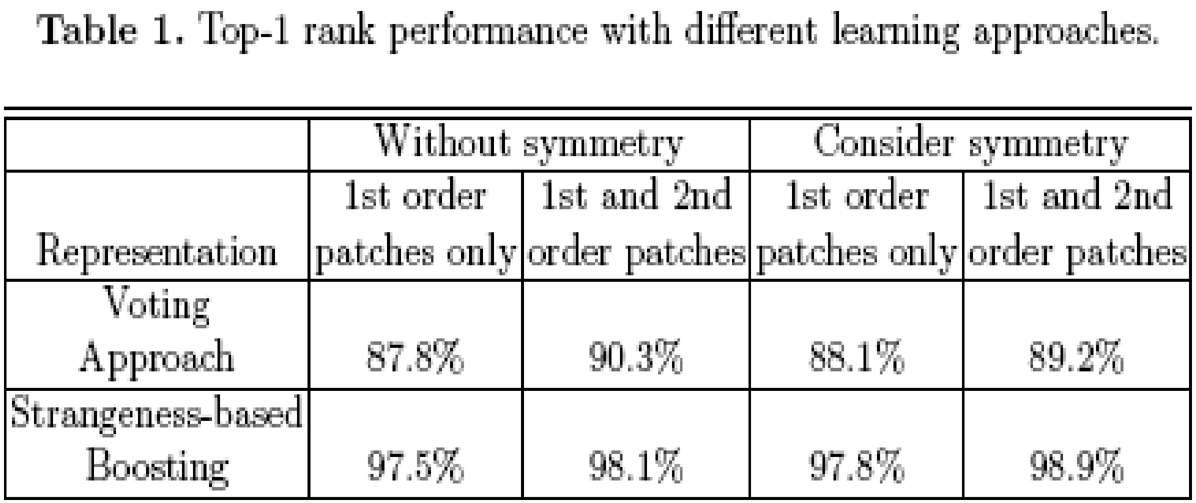
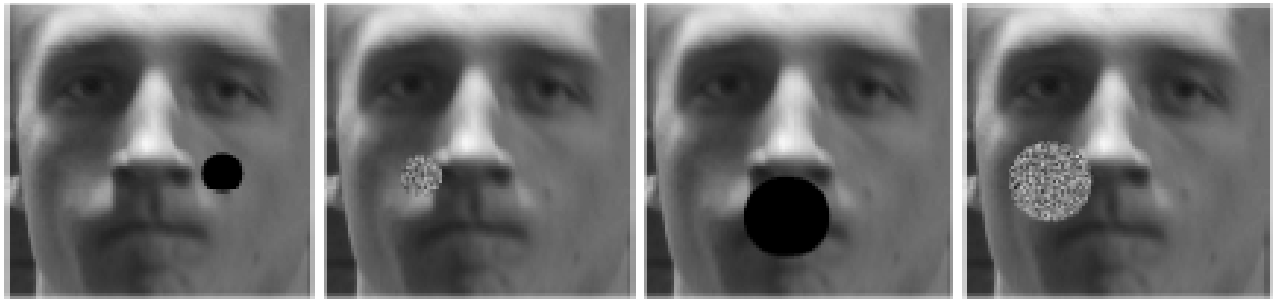


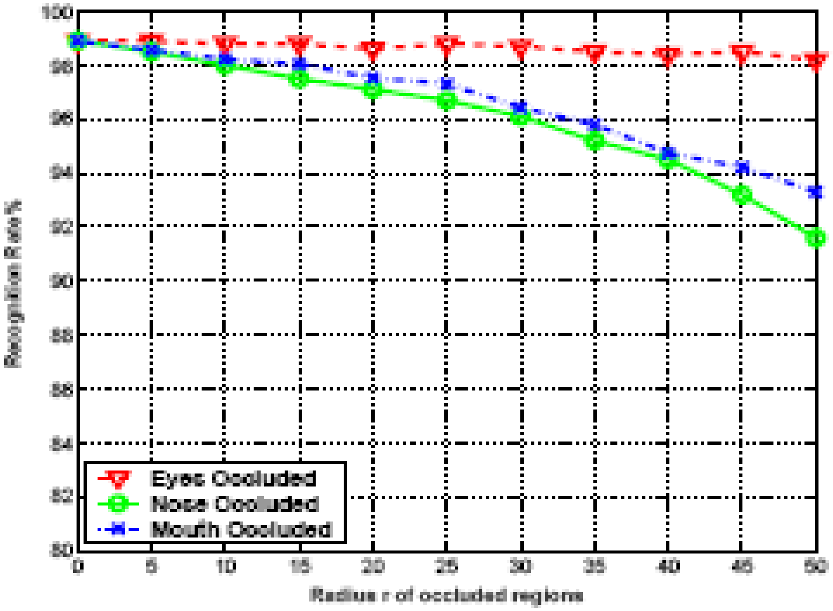
Table 1. Top-1 rank performance with different learning approaches.

	Without symmetry		Consider symmetry	
Representation	1st order patches only	1st and 2nd order patches	1st order patches only	1st and 2nd order patches
Voting Approach	87.8%	90.3%	88.1%	89.2%
Strangeness-based Boosting	97.5%	98.1%	97.8%	98.9%

Occluded Face Images



Recognition Rates when Eye, Nose or Mouth is Occluded



R&D

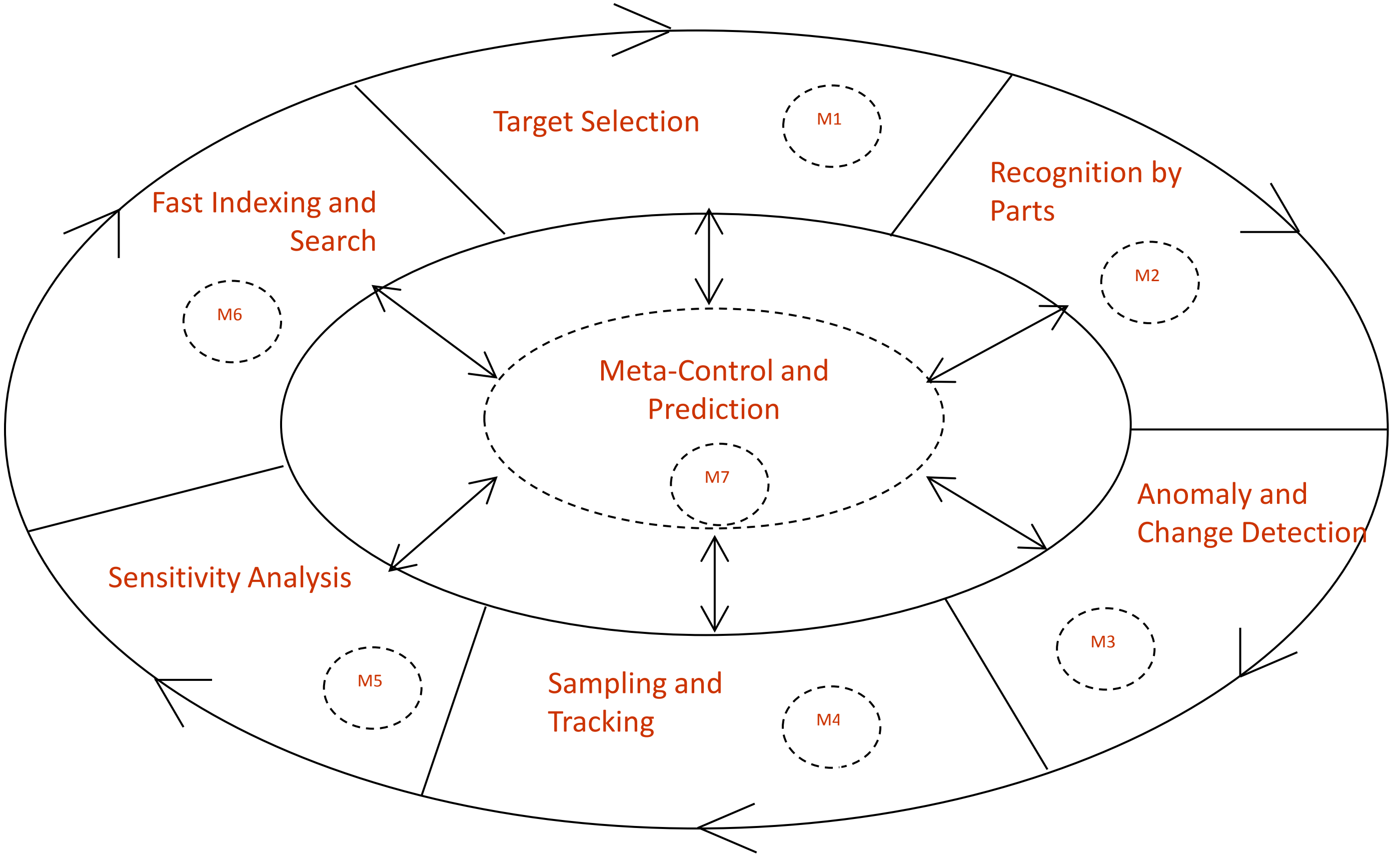
W5+ (*what, who, where, when, why, how*)

- Active Learning (QBT) [PAMI 2008]
- Adaptive and Robust Correlation Filters (ARCF) [CVIU 2008]
- Anthropometrics / Soft Biometrics and Context [PRL 2010]
- Change Detection Using Martingale [PAMI 2010]
- Data Fusion [IJPRAI 2009]
- Interoperability ~ Evidence-Based Management (EBM) [BIOMS 2012]
- Lapsed-Time (Aging) FR Using Multi-Task (Transfer) Learning and Covariate Shift (2013 -)
- Re-Identification [PRL 2012]
- Sensitivity / Stability Analysis ~ Generalization, Prediction, and Revision ~ Adversarial Learning and Consensus (2013 -)
- Reverse Learning and CBIR (2013 -)

Change Detection and Martingale

- Skewness (a measure of the degree of asymmetry of a distribution), deviates from close to zero (for uniformly distributed p-values) when a model change occurs. The skewness is small and stable when there is no change. The skewed p-value distribution plays an important role in the martingale test for change detection as small p-values inflate martingale values. When the observed data points are no longer exchangeable, due to change in the underlying distribution, the p-values are also no longer uniformly distributed over [0, 1]. In particular, the p-values have smaller value, due to the fact that newly observed data points are likely to have higher strangeness values compared to the previously observed data points.

Re-Identification Using Evidence-Based Management



Protocols and Validation
-- Curb Your Enthusiasm --

- Illusion of Progress
- Reliability and Robustness
- Biometric DB ~ **real-world image variation** (Pinto, Cox, and Di Carlo (2008)
- Face Aging
- Biometric Menagerie and Error Analysis
- Best Practices and Protocols

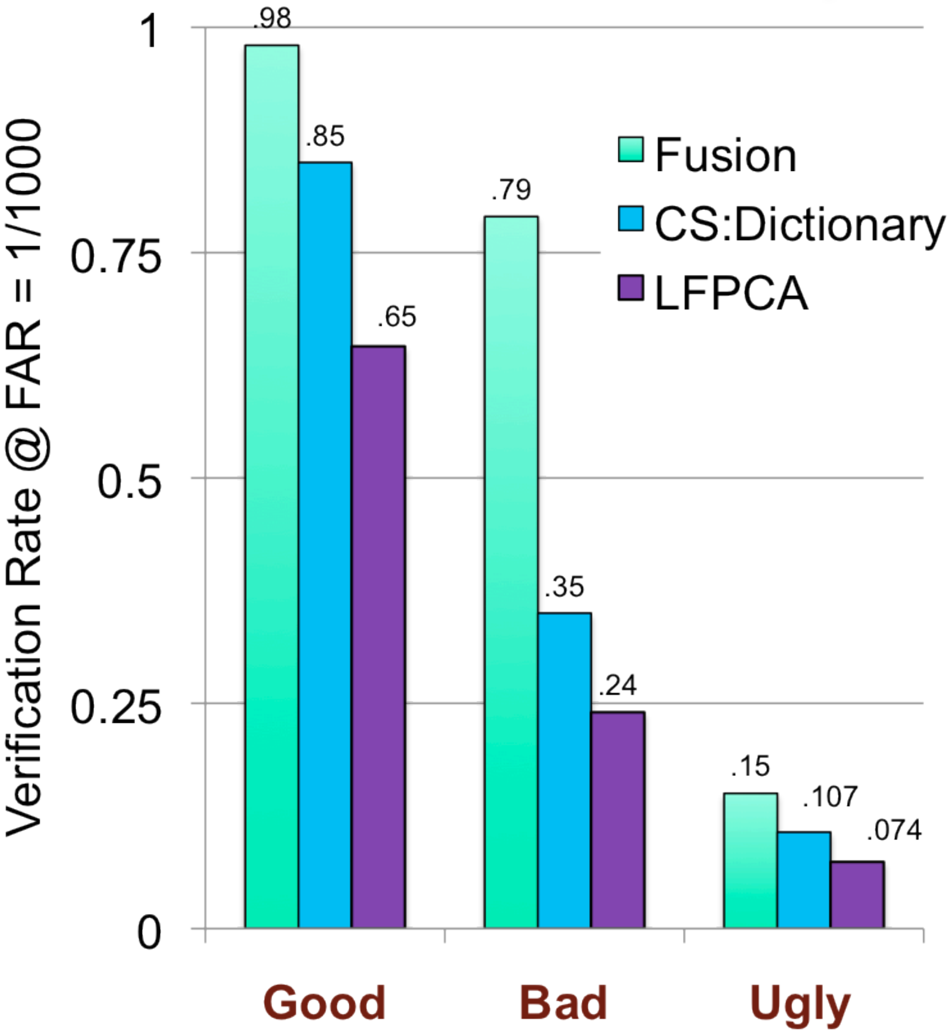
Illusion of Progress and Marginal Improvements
- Hand (2006): Interpretability and Parsimony -

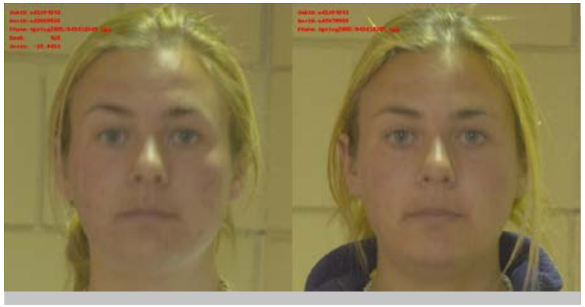
- ***Flat Maximum*** effect (model fitting and progressive refinement): Large gains in predictive accuracy are won using relatively simple models at the start of the process.
- ***Sample Selectivity Bias*** and ***Population Drift***: In many, perhaps most, real classification problems the data points in the design are not, in fact, randomly drawn from the same distribution as the data points to which the classifier will be applied.
- ***Problem Uncertainty***: Density, margin, unbalanced populations, and error in labels
- ***Risk Analysis***: Arbitrary assumptions, optimization criteria, choices made ("over fitting"), and relative costs ("losses") for different kinds of misclassification.
- ***Evaluation***: "Laboratory" conditions may not transfer to real-world conditions.
- ***Base Rate Fallacy:*** Prevalence and Precision

Validation

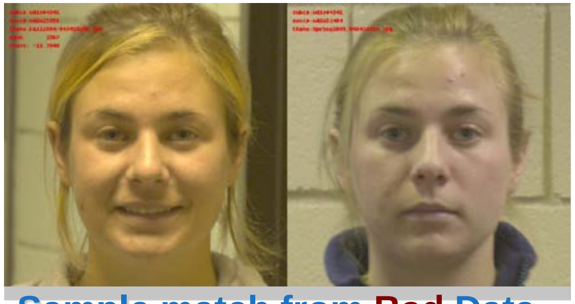
“Yet there also comes a time when performance on a benchmark reaches ceiling performance or methods become over-engineered for nuances of a data set, and modest performance gains may be indicative of over-fitting. Alternatively, some new works or operational scenarios may push the envelope in directions that are not well represented with existing benchmarks; in such cases, authors may need to develop alternative benchmarks and justify this need in subsequent publications. Interestingly, real world face recognition methods that achieve state-of-the-art performance on data sets like Learning from the Wild (LFW) may actually perform worse on constrained, frontal data sets like FERET. We should not be surprised by this, and we should embrace methods for where they are effective” (Hua et al., 2011)

The “Good, Bad, Ugly”

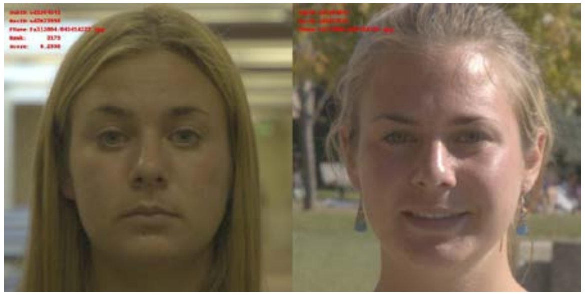




Sample match from Good Data



Sample match from Bad Data (challenging)



Sample match from Ugly Data (very challenging)

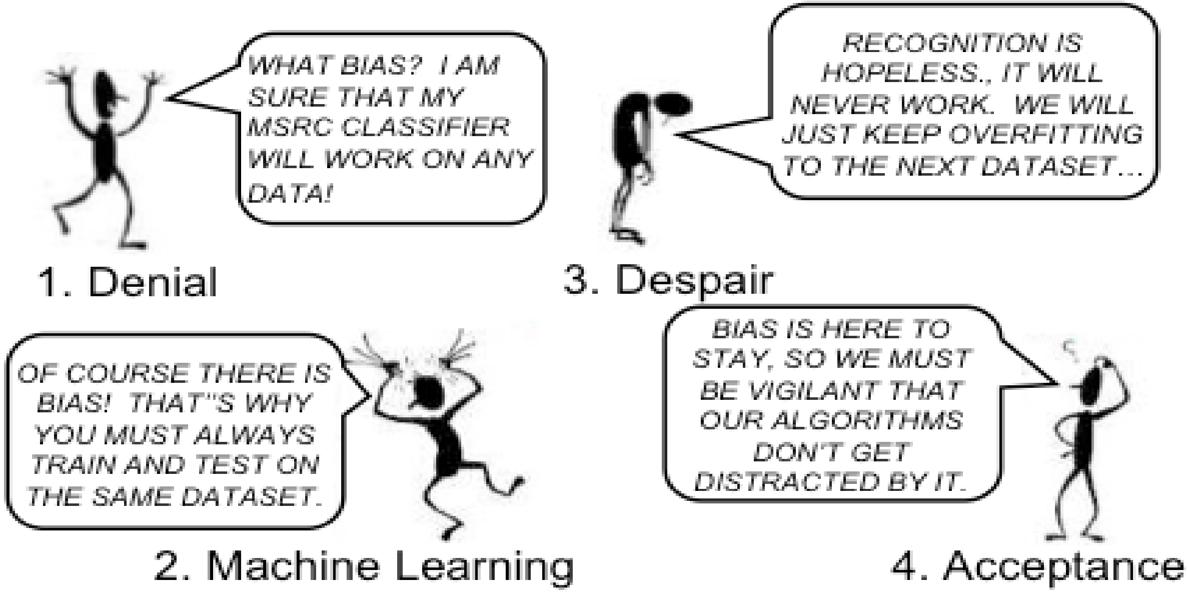
A Moral for Learning Machines

- AI Expert Newsletter (March 2006) -

- *Life* magazine once ran a two-page spread of about 40 photographs of different persons. Half of them were of college professors, scientists, and esteemed businessmen. The other half were criminals ranging from thieves to rapists to murderers. The magazine feature was a fun contest for the reader to see if he could tell the good citizens from the criminals. My wife and I tried it. My score was about 30 percent right; her score was 100 percent right. Did she have special insight? Yes, but not about faces. She observed that half the photographs had the same draped background and deduced correctly that the criminals were photographed at the same locale.
- This story comes from *How to Draw Caricatures* by Lenn Redman. It reminds me of the urban legend about the neural net which was trained to spot camouflaged tanks. Trained on 100 battlefield scenes, each of which contained either a tree with a tank hiding behind it, or a tree with no tank behind it, the net did indeed learn to sort scenes with tanks from scenes with no tanks. Not as a result of the tanks; but because the images with tanks had been taken on a cloudy day while images without tanks had been taken on a sunny day.

Datasets and Interoperability
-Unbiased Look at Dataset Bias -

http://people.csail.mit.edu/torralba/research/bias/

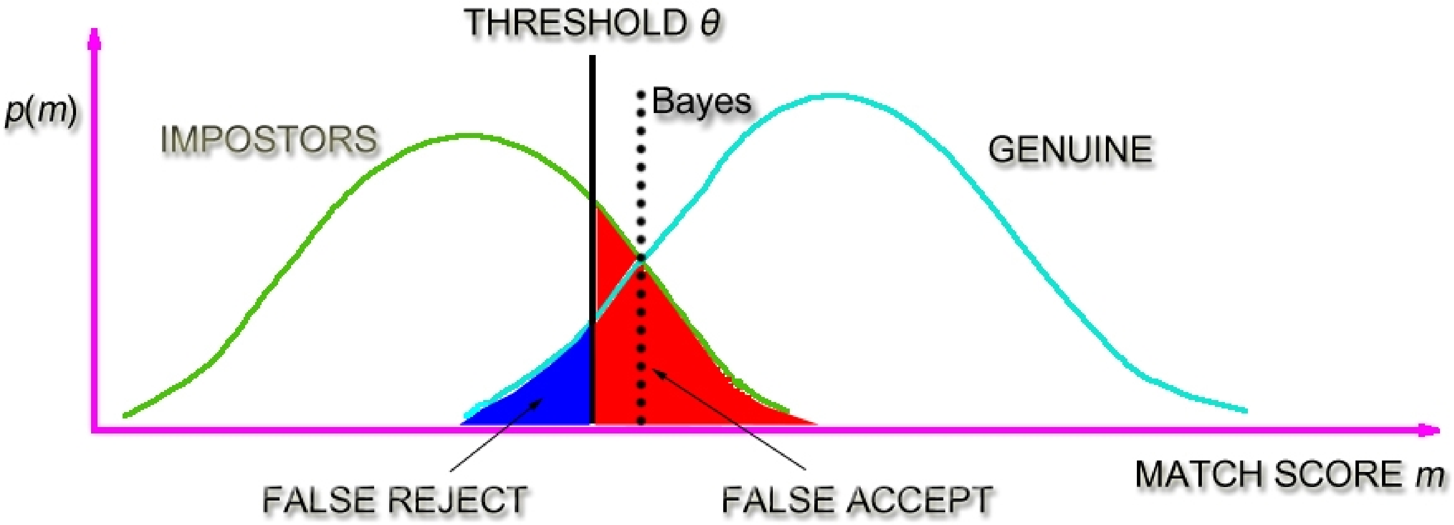

WHAT BIAS? I AM SURE THAT MY MSRC CLASSIFIER WILL WORK ON ANY DATA!
1. Denial
OF COURSE THERE IS BIAS! THAT''S WHY YOU MUST ALWAYS TRAIN AND TEST ON THE SAME DATASET.
2. Machine Learning
RECOGNITION IS HOPELESS., IT WILL NEVER WORK. WE WILL JUST KEEP OVERFITTING TO THE NEXT DATASET…
3. Despair
BIAS IS HERE TO STAY, SO WE MUST BE VIGILANT THAT OUR ALGORITHMS DON'T GET DISTRACTED BY IT.
4. Acceptance

Face Aging

- Alignment, Correspondence, and Registration ~ CONSENSUS
- Age Estimation
- **Covariate Shift** -- [1] $P_T(x) \neq P_t(x)$ & $P_T(y|x) = P_t(y|x)$; [2] Dataset Bias ~ [**1**] Importance Sampling and Importance Weighting; [**2**] Database Collection and Demographics; [**3**] Region-Based Approach; [**4**] Soft Biometrics; and [**5**] Transfer Learning

☺ CHALLENGE ☺

CHALLENGE
- setting thresholds -



Quo Vadis

- Evidence Accumulation and Managerial Systems
- [1] Shape, Texture, and Change / Dynamics; and [2] Parts and Topics
- Recognition and Tracking
- [3] Learning with Side Information (Anchors and Transformations); and [4] Consensus and Local and Global Consistency (Semi-Supervised Learning)

What is permanent and unique is the change itself.

Heraclitus of Ephesus (c. 500 B.C.) claimed that all things are in flux and that everything is constantly changing. One can not step into the same river twice since the river is never the same. Hence the variability of biometrics and the “permanent” challenge to handle in a reliable fashion the **ever changing human faces**. Parmenides of Elea (c. 515 – 450 B.C.) thought quite differently from Heraclitus. He sought what is permanent and never changing, and proposed a duality of appearance and reality. The changing world registered by our senses is merely an illusion. There are alternatives or hypotheses about illusory appearances and they have to be searched to ferret out the reality behind them. Parmenides claimed that it is only **through reason** that one can **indirectly learn about “real” existence**, which by itself is **permanent**, i.e., unchanging and unmoving. How to navigate between those seemingly Scylla and Charybdis rocks of beliefs? It was Democritus (c. 460 – 370 B.C.) who, while trying to reconcile between Heraclitus and Parmenides, came to claim that there is place for both **permanence** and **change**.

What is permanent and unique is the change itself.

Permanence is found in the essence of things, while change comes from motion. According to David Hume there appear to be only three principles of connection among ideas, namely resemblance, contiguity (in time or place), and cause or effect. This corresponds to similarity across the face space, spatial-temporal coherence, and learning, inference, and reasoning. There is a growing realization that tracking and recognition, i.e., change and permanence, are complementary to each other. What is indeed unique to objects, in general, and faces, in particular, and constitutes their essence, is the particular way each human face changes or morphs across multiple views and time**.** The spatial – temporal trajectories traced by faces are unique to each individual and should serve for their reliable identification and authentication notwithstanding uncontrolled settings.

Media in the Wild

“change is unique”

THANKS!