Face Recognition:
from Biological to Artificial Deep Networks

Massimo Tistarelli

Computer Vision Laboratory

University of Sassari – Italy

tista@uniss.it

Credits

- From the laboratory staff:

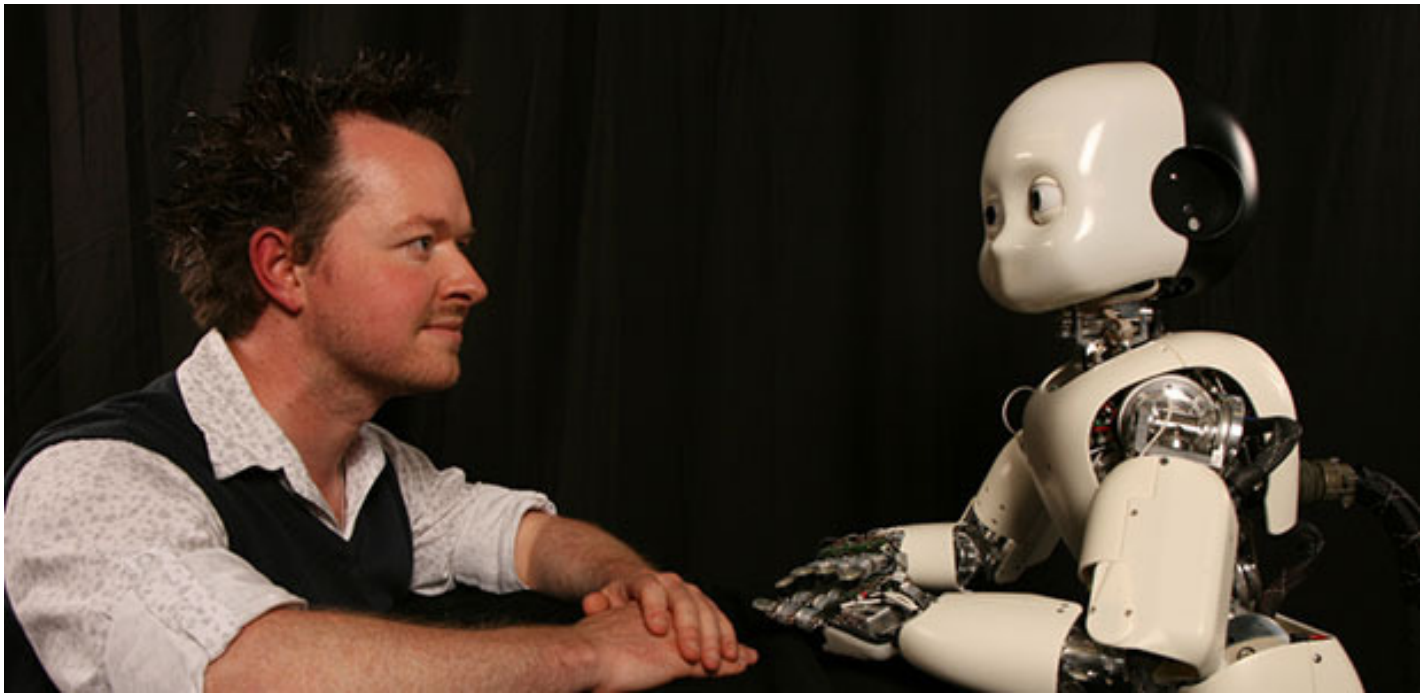
Linda Brodo
Marinella Cadoni
Filippo Casu
Massimo Gessa
Enrico Grosso
Souad Khellat Khiel
Andrea Lagorio
Ludovica Lorusso
Gianluca Masala
Seth Nixon
Ajita Rattani
Elif Surer
Yunlian Sun
Humera Tariq
Norman Poh (past visiting)
Daksha Yadav (past visiting)
Yu Guan (past visiting)
Marcos Ortega Hortas (past visiting)
Albert Ali Salah (past visiting)

Credits

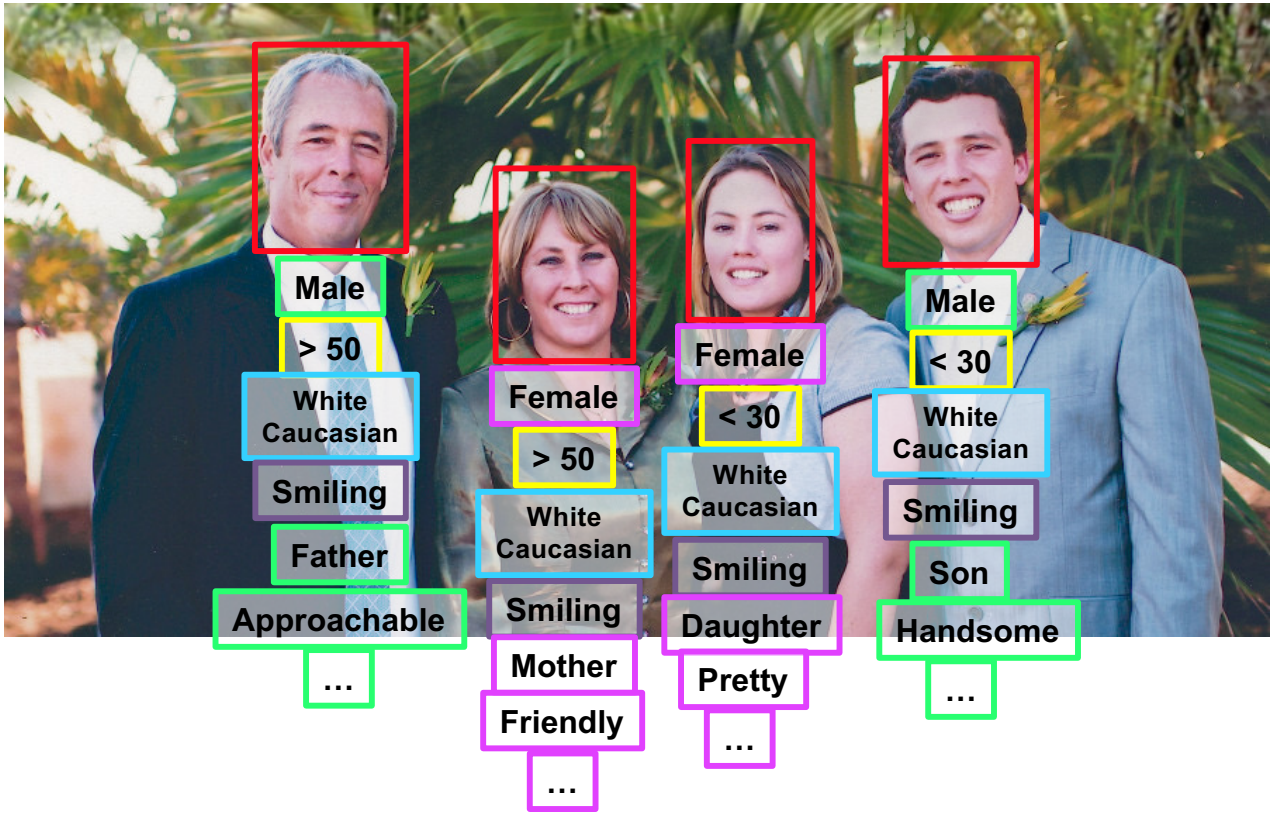
- …and other labs:

Manuele Bicego – University of Verona
Rama Chellappa – University of Maryland
Anil Jain – Michigan State University
Alice O'Toole – University of Texas at Dallas
Chang-Tsun Li – University of Warwick
Jonathon Phillips – NIST
Norman Poh – University of Surrey

Natural vs Artificial Intelligence

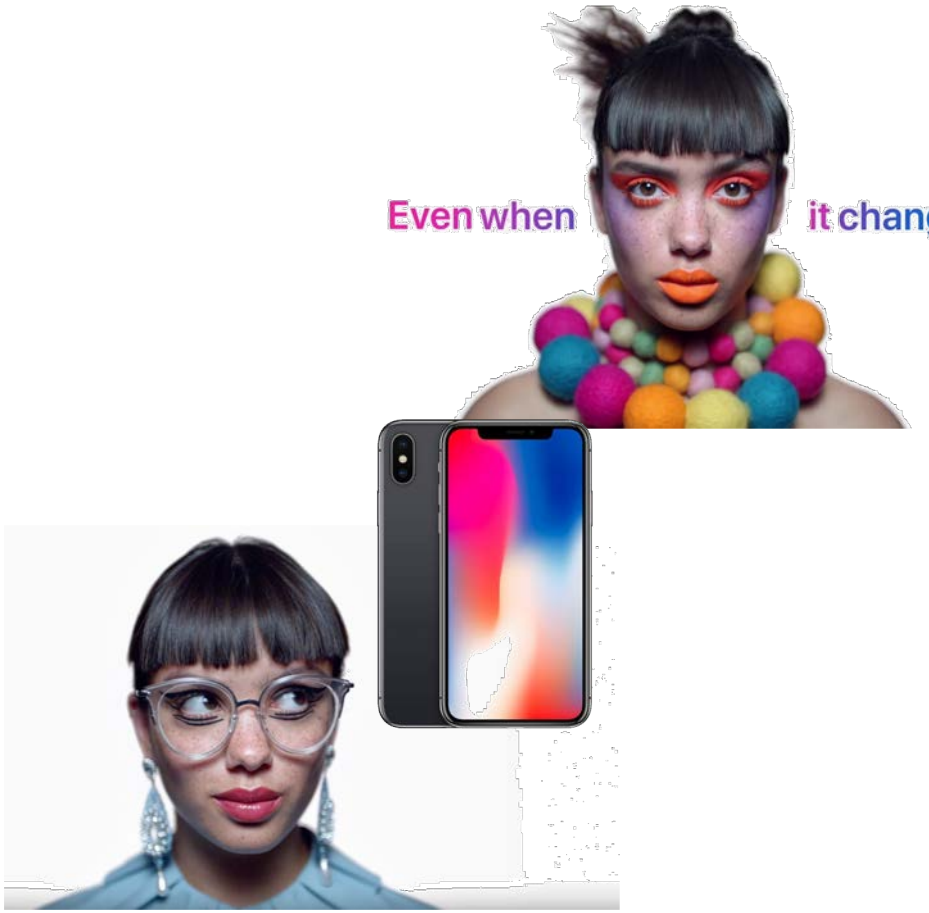


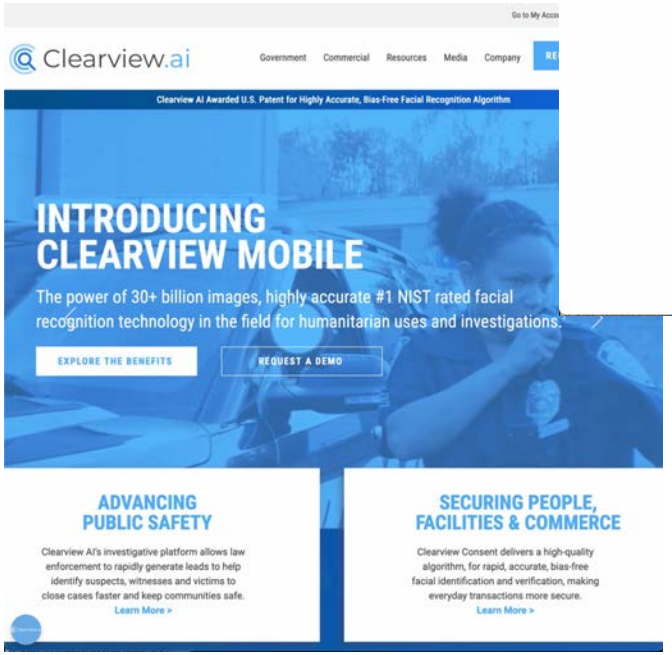
One image… A lot of information

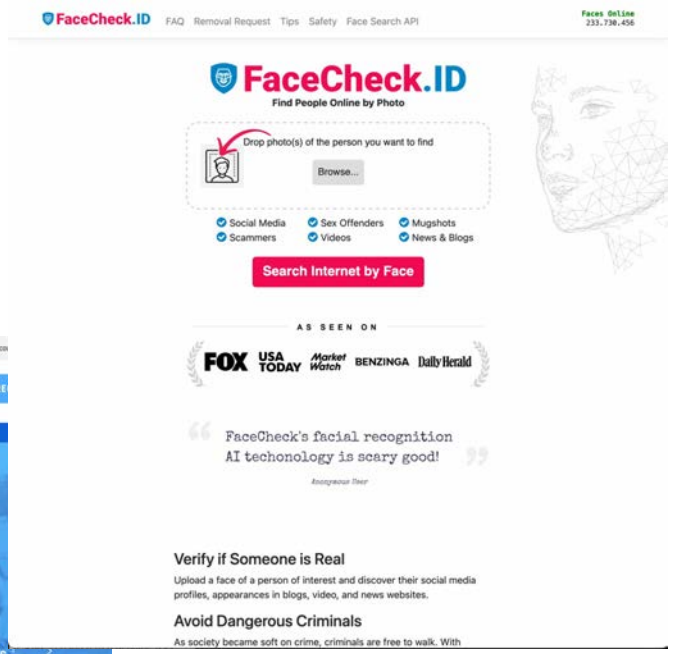


Too many faces!

Faces in the market

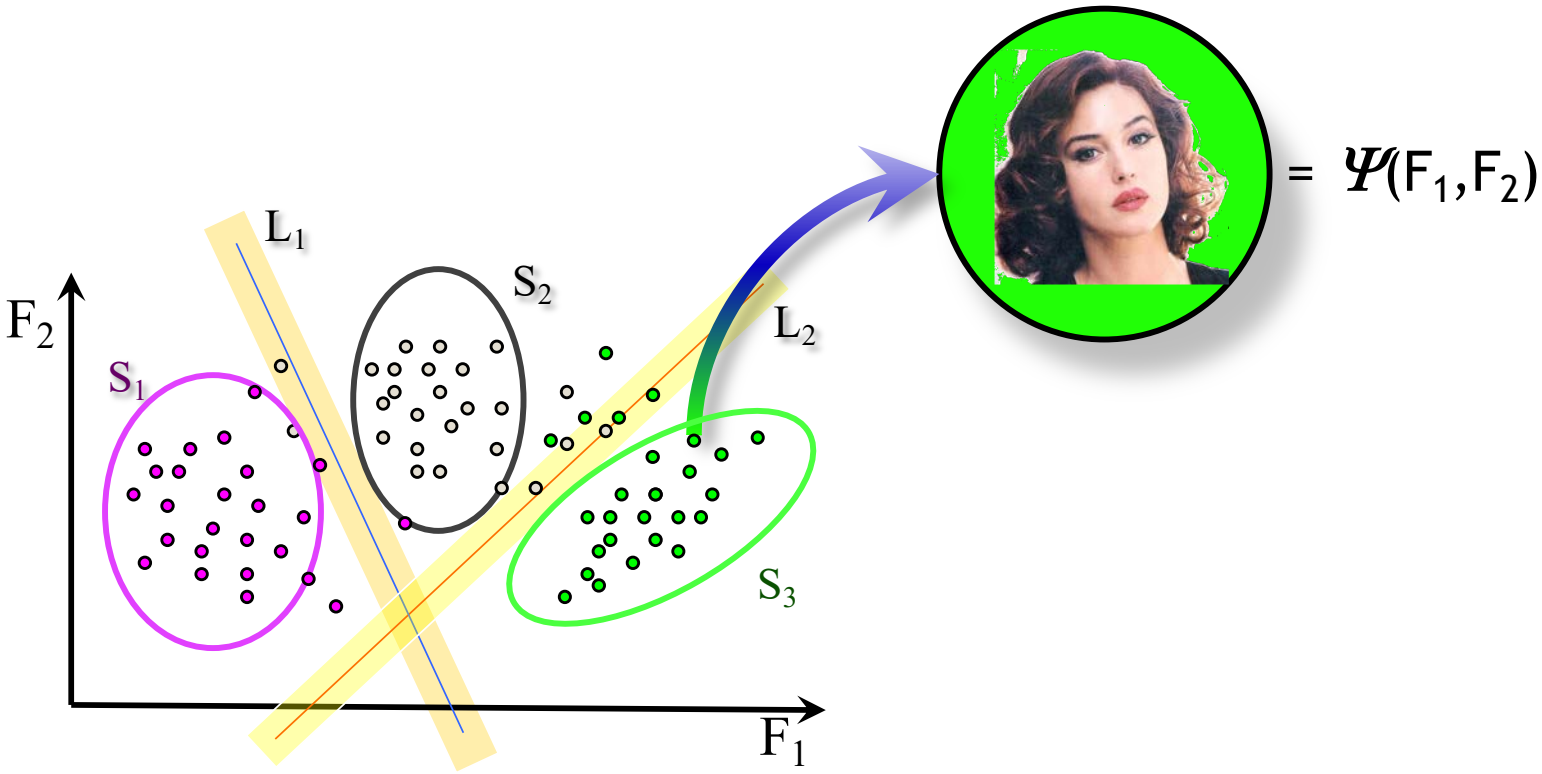






Face Recognition

A class (***identity***) separation problem



Genuine and Impostor scores

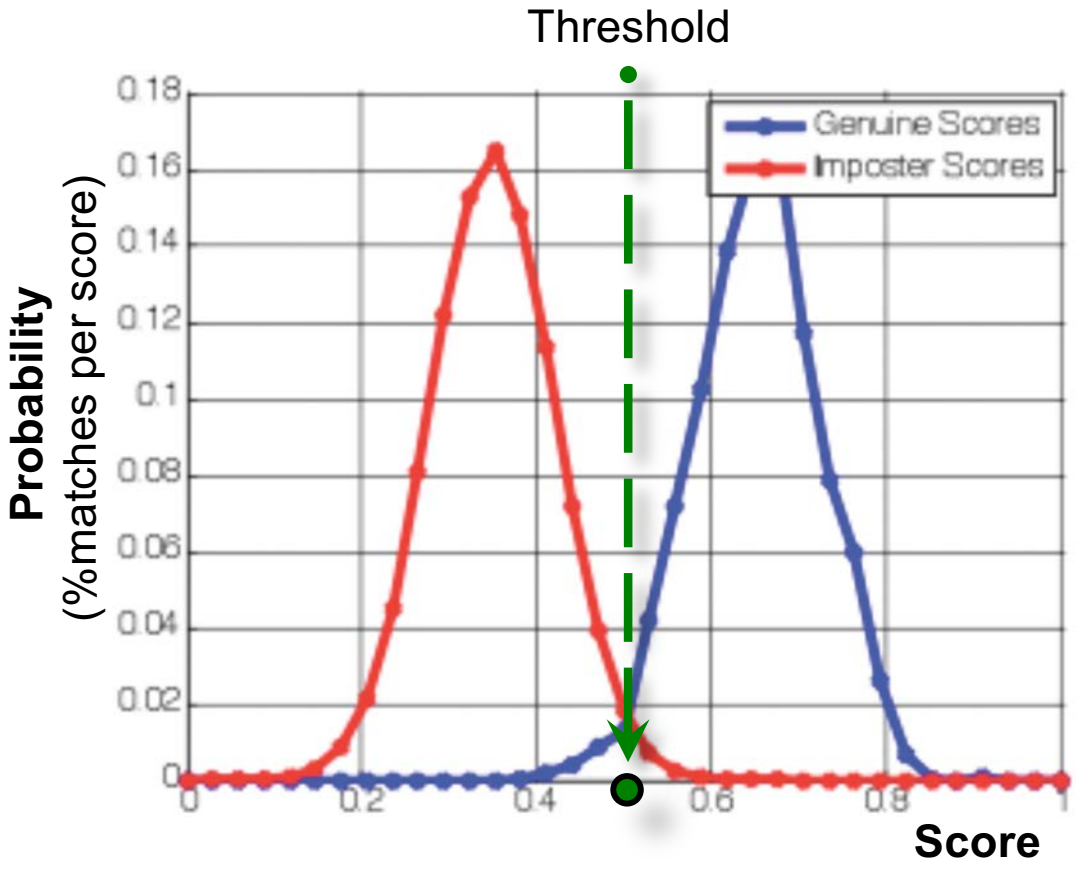
- **Genuine score**: ***Match score*** (*degree of similarity* or *closeness*) computed by comparing two biometric samples from the **same** individual.

- **Impostor score**: ***Match score*** computed by comparing two biometric samples originating from **different** individuals.

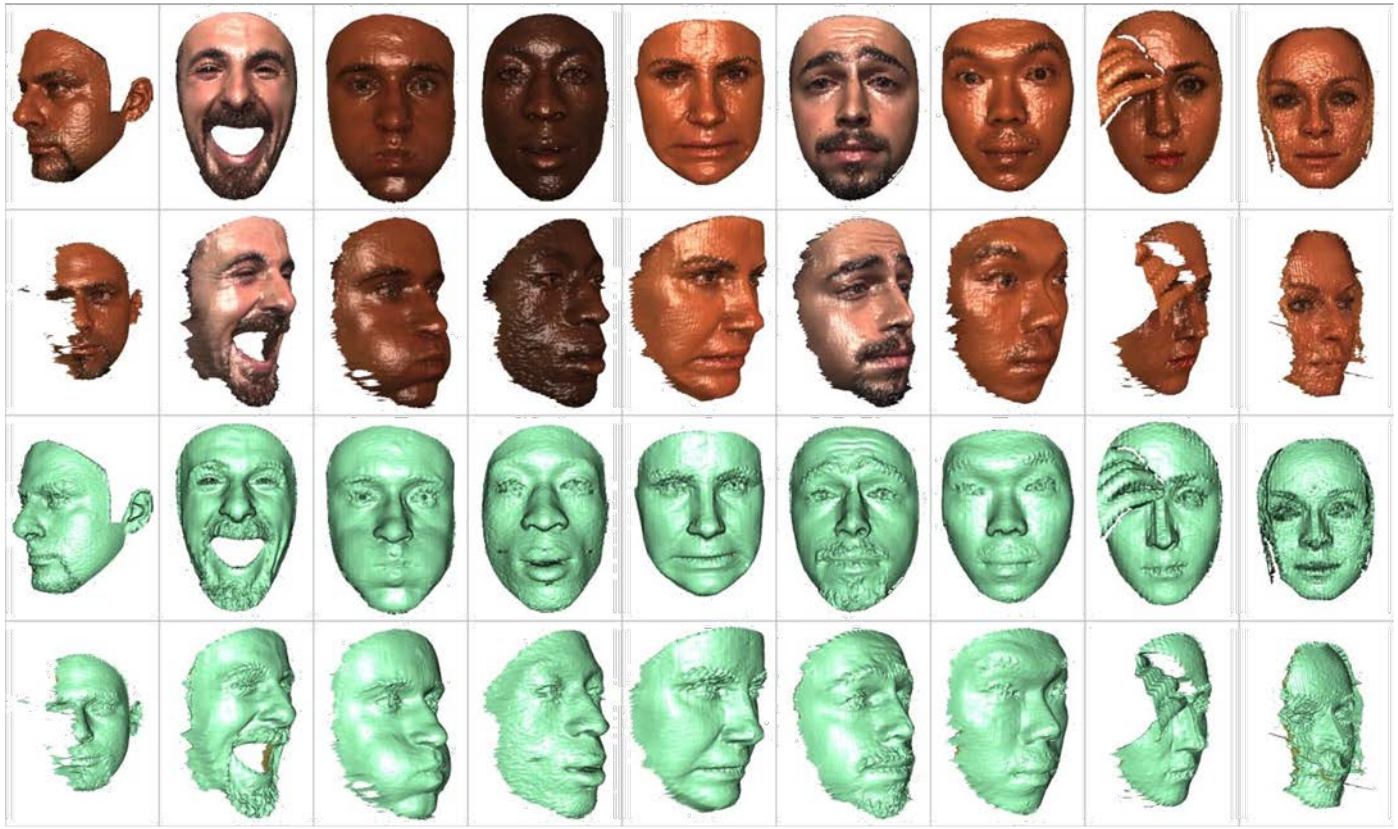
  Therefore, **a genuine user score should be always greater than an impostor score**.

- A **threshold** (or **classifier**) is used to determine if a score is related to a genuine user or an impostor.

Match score distributions



Face shape and texture



A. Savran, N. Alyüz, H. Dibeklioğlu, O. Çeliktutan, B. Gökberk, B. Sankur, L. Akarun, "**Bosphorus Database for 3D Face Analysis**", The First COST 2101 Workshop on Biometrics and Identity Management (BIOID 2008) Roskilde University, Denmark, May 2008.

Visual challenges

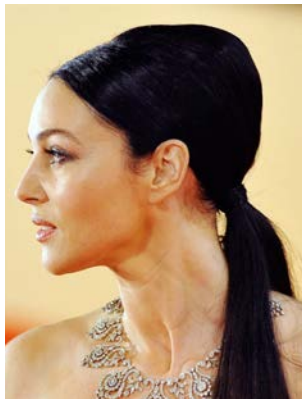
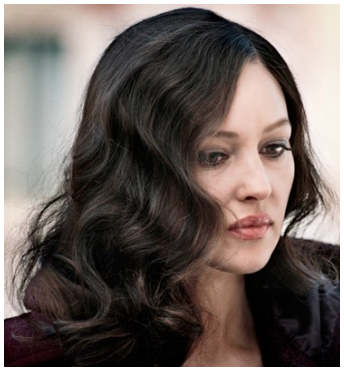
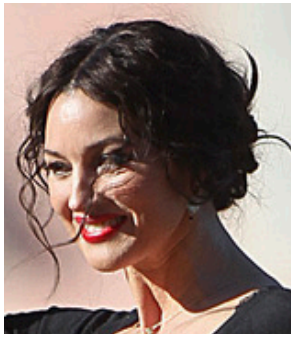
A – Aging

P – Pose

I – Illumination

E - Expression



Visual challenges

Esthetic surgery

Make up

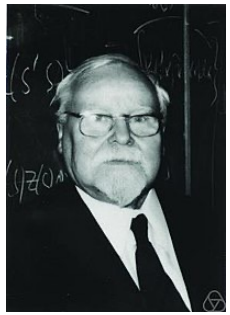
Visual challenges

UMD-AA Mobile Device Database

U. Mahbub, S. Sarkar, V. M. Patel and R. Chellappa, "**Active user authentication for smartphones: A challenge data set and benchmark results**," 2016 IEEE 8th International Conference on Biometrics Theory, Applications and Systems (BTAS), Niagara Falls, NY, 2016, pp. 1-8..

An ill-posed problem

Jacques Hadamard



Andrej Tikhonov

Two adverse conditions:

1) **Noise** in the data (many sources, including **A.P.I.E.**)
2) **Dimensionality** of the data (from 4D to 2D)

Solution: Regularization

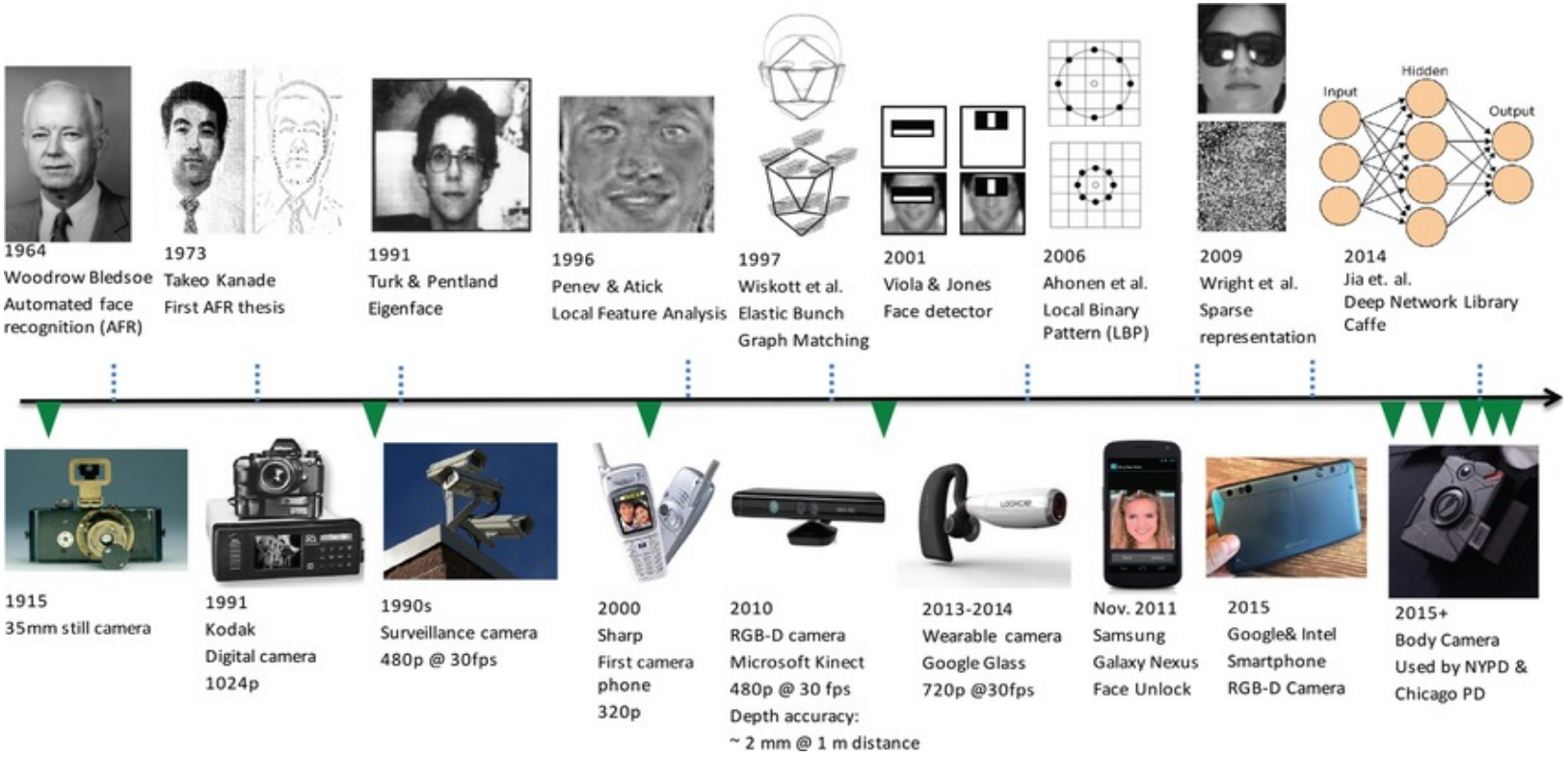
J. Hadamard, "**Sur les problemes aux derivees partielles et leur signification physique**". In: Princeton University Bulletin, 1902, 49–52.

A.N. Tikhonov, "**On the stability of inverse problems**". Doklady Acad. Sci. USSR 39 (1943), 176–179.

A.N. Tikhonov, "**On the solution of ill-posed problems and the method of regularization**". Dokl. Akad. Nauk SSSR 151(3) (1963), 501–4.

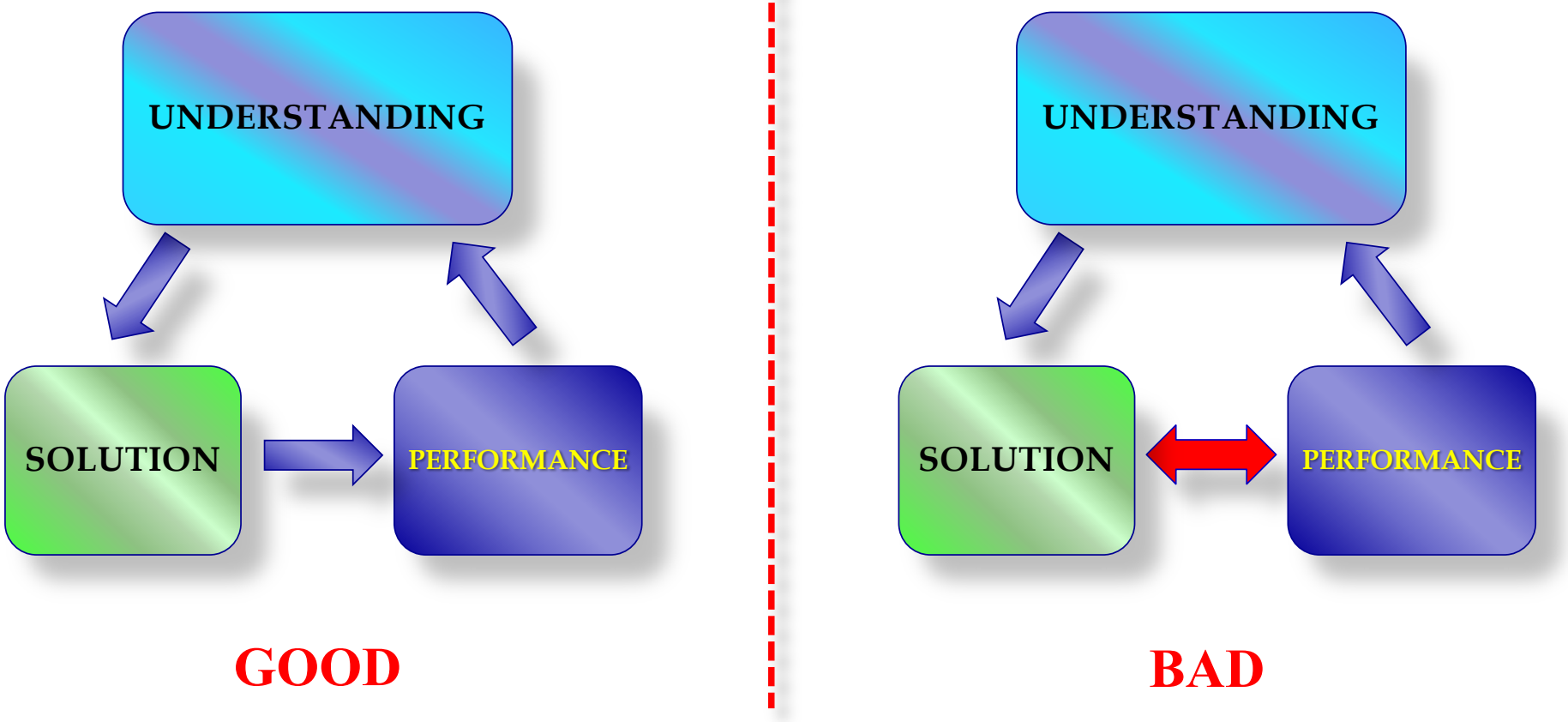
A. N. Tikhonov and V. Ya. Arsenin, "**Solutions of Ill-Posed Problems**". Wiley, New York, 1977.

Face recognition milestones



A. **Jain, K, Nandakumar, A. Ross**, "50 Years of Biometric Research: Accomplishments, Challenges, and Opportunities", Pattern Recognition Letters 79:80-105, 2016.

Good research or bad research?



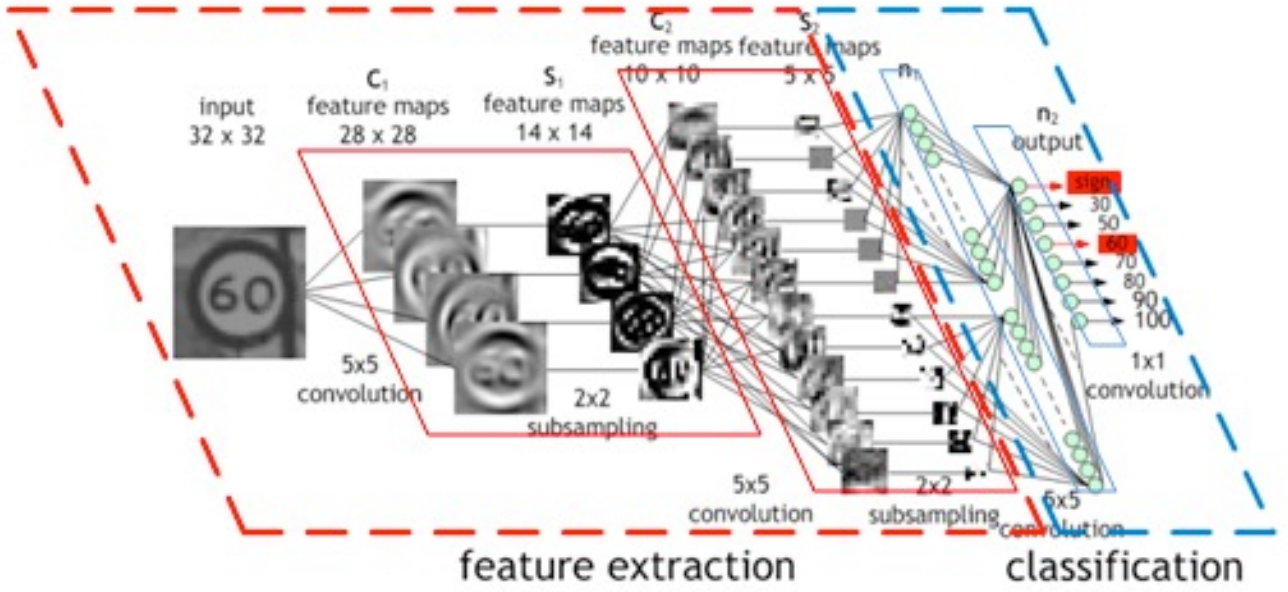
Common mistakes

1. Start **programming** before **thinking**.
2. Building a system **blindly** combining a number of already available algorithms.
3. Performing **blind** **tests** with available tools and datasets *(«Quick prototyping»*?).
4. Twickling the **parameters** until you obtain the **desired performance**.
5. Arbitrarily **selecting the data** from the available datasets **after** performing the initial testing.
6. Making **strong statements** without a solid proof.
7. Making **unrealistic assumptions**.

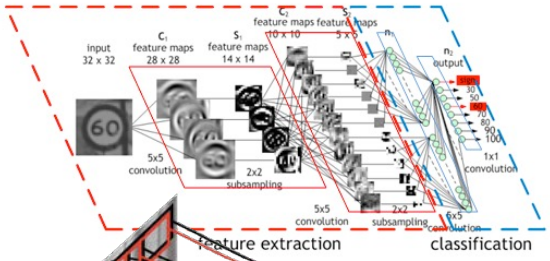
Addressing the problem

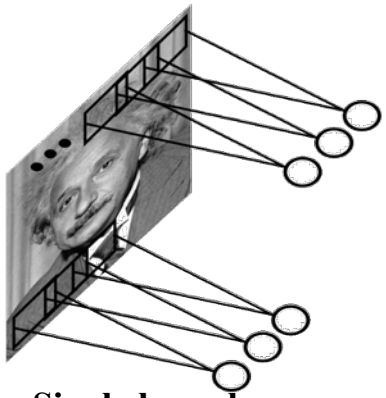
1. Analyze the **problem**, the available **data** and the **constraints**.
2. Make a **bibliographical search** (don't try to re-invent the wheel… one is enough).
3. Define a **model** describing the **physics** of the **event.**
4. Find a **mathematical framework** which may bring to a solution.
5. Carefully **design** an **experimental set-up**.
6. Collect or acquire a **statistically meaningful dataset**.
7. Start **programming**.
8. Perform an **evaluation test** to define the **parameters space**.
9. Start testing and collecting results, especially the **failing modes**.
10. Perform a **comparative analysis** of the results with other approaches at the ***current*** state of the art.
11. **Go back to item 3**.

Convolutional Neural Networks

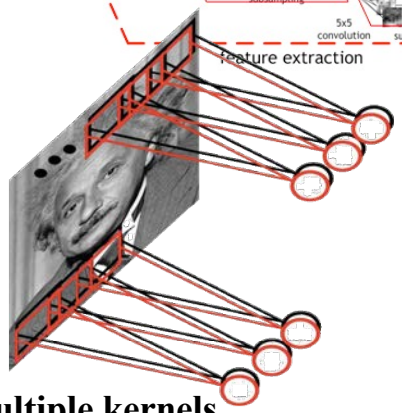


Convolutional Neural Networks

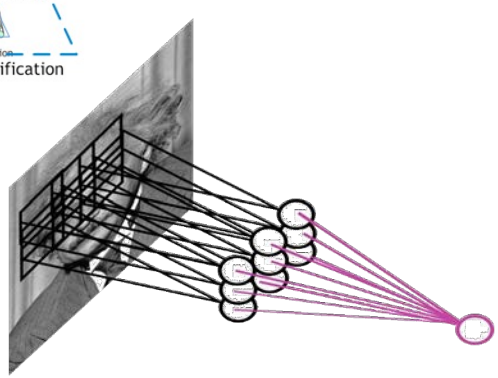




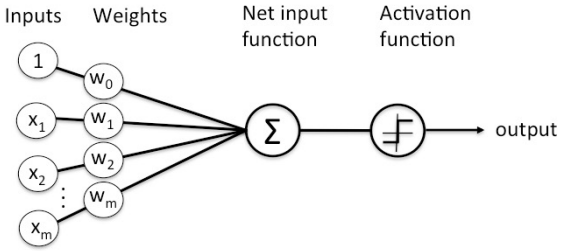
Single kernel Convolution



Multiple kernels Convolution



Spatial Pooling

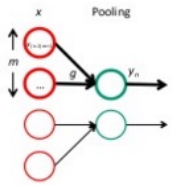


Let m be the size of pooling region, x be the input, and y be the output of the pooling layer. subsample$(f, g)[n]$ denotes the n-th element of subsample(f, g).

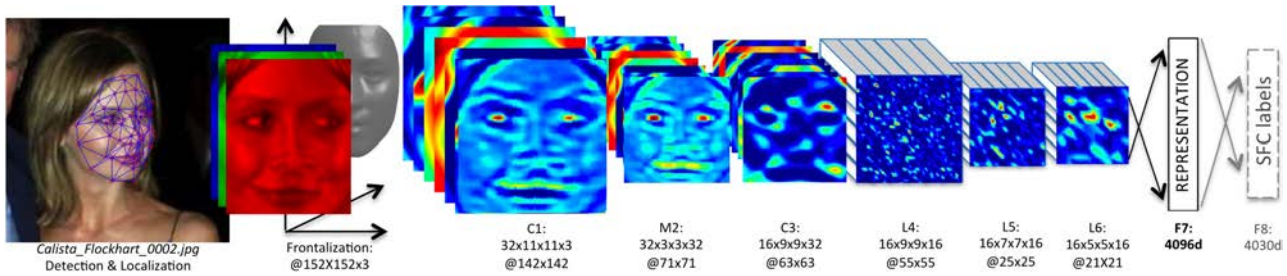
$$y_n = \text{subsample}(x, g)[n] = g\left(x_{(n-1)m+1:nm}\right)$$

$$y = \text{subsample}(x, g) = [y_n]$$

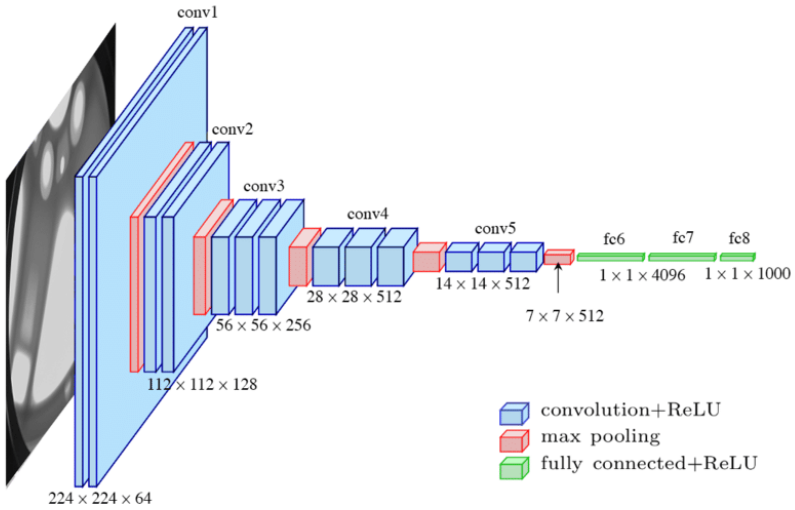
$$g(x) = \begin{cases} \dfrac{\sum_{k=1}^{m} x_k}{m}, \dfrac{\partial g}{\partial x} = \dfrac{\mathbf{1}}{m} & \text{mean pooling} \\ \max(x), \dfrac{\partial g}{\partial x_i} = \begin{cases} 1 \text{ if } x_i = \max(x) \\ 0 \text{ otherwise} \end{cases} & \text{max pooling} \\ \|x\|_p = \left(\sum_{k=1}^{m} |x_k|^p\right)^{1/p}, \dfrac{\partial g}{\partial x_i} = \left(\sum_{k=1}^{m} |x_k|^p\right)^{1/p-1} |x_i|^{p-1} & \text{L}^p \text{ pooling} \\ \text{or any other differentiable } \mathbf{R}^m \to \mathbf{R} \text{ functions} \end{cases}$$



Convolutional Neural Networks

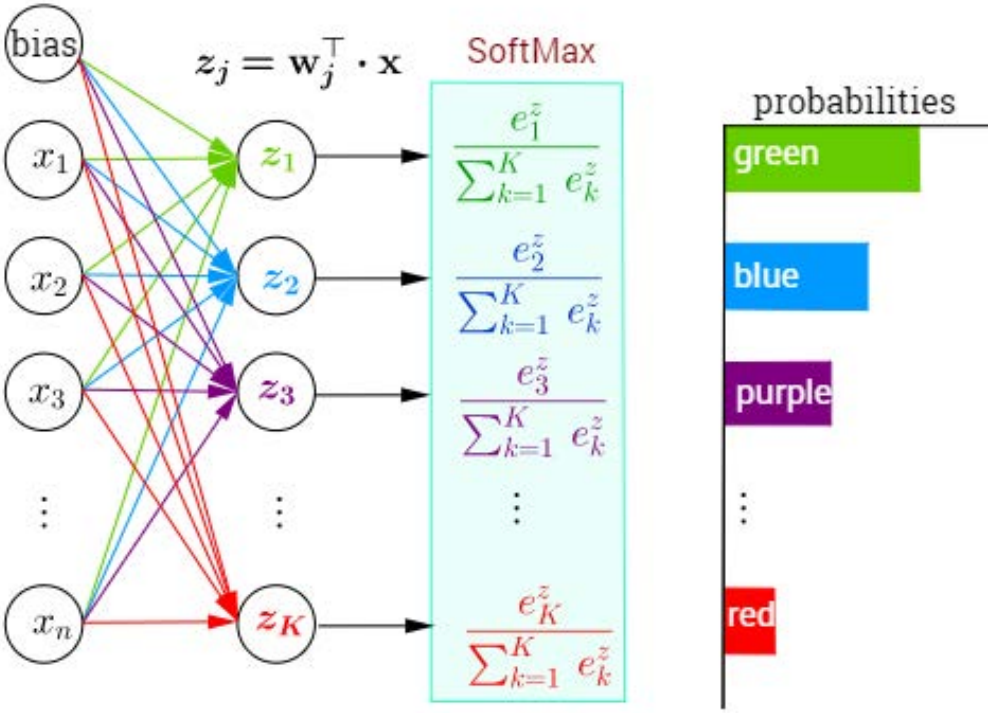


Y. Taigman, M. Yang, M. Ranzato, L. Wolf "**DeepFace: Closing the gap to human-level performance in face verification**" CVPR, 2014



O. M. Parkhi, A. Vedaldi, A. Zisserman "**Deep Face Recognition**" British Machine Vision Conference, 2015

Convolutional Neural Networks




```
def softmax(X):
        exps = np.exp(X)
        return exps / np.sum(exps)
```

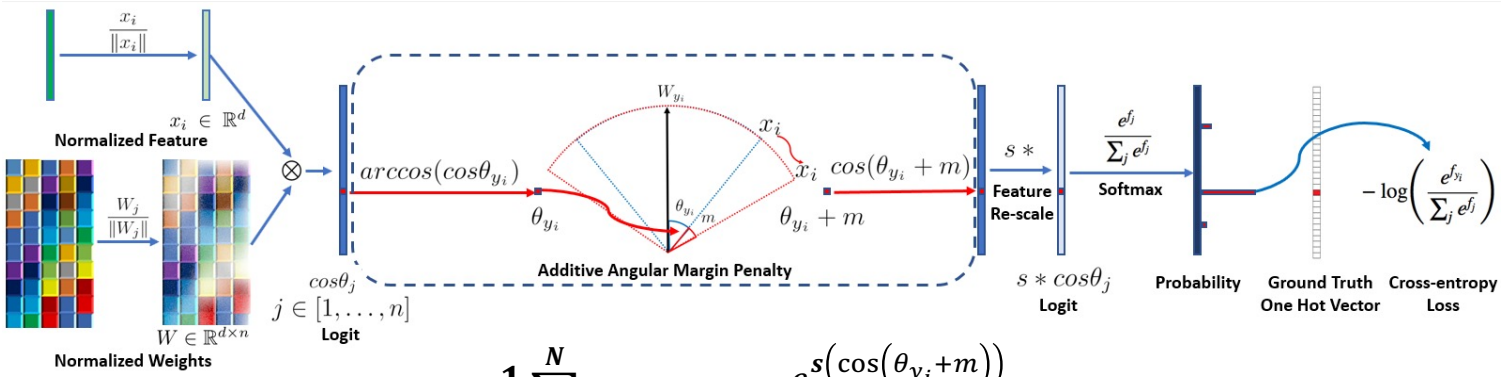
Loss functions

Cross entropy indicates the distance between what the model **believes** the output distribution should be, and what the original distribution **really is**:

$$\boldsymbol{H(y,p) = -\sum_i y_i \log(p_i)}$$

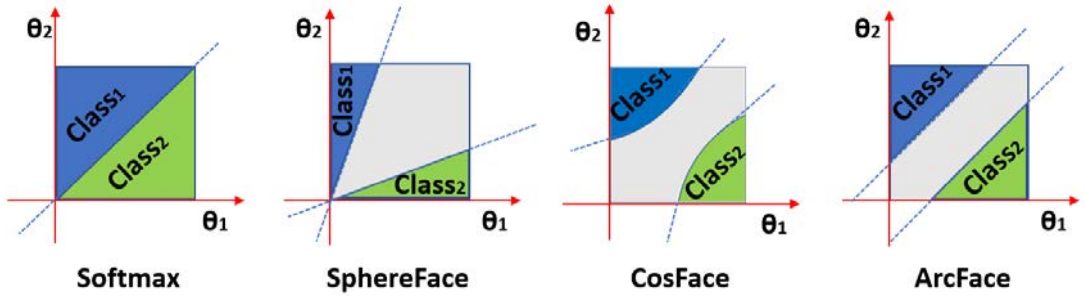
```
def cross_entropy(X,y):

""" X is the output from a fully connected layer (num_examples x num_classes)
y is labels (num_examples x 1)
Note that y is not one-hot encoded vector. It can be computed as y.argmax(axis=1) from one-hot encoded
vectors of labels if required.
"""
    m = y.shape[0]          # We use multidimensional array indexing to extract
    p = softmax(X)          # softmax probability of the correct label for each sample.

    log_likelihood = -np.log(p[range(m),y])
    loss = np.sum(log_likelihood) / m
    return loss
```

Loss functions



$$L(s) = -\frac{1}{N}\sum_{i=1}^{N} \log \frac{e^{s\left(\cos\left(\theta_{y_i}+m\right)\right)}}{e^{s\left(\cos\left(\theta_{y_i}+m\right)\right)} + \sum_{j=1, j\neq y_i}^{n} e^{s\cos\theta_j}}$$

θ_j is the angle between the weight W_j and the feature x_i ; $s = \|x_i\|$



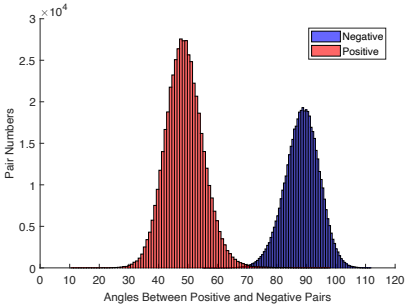
Deng J, Guo J, Yang J, Xue N, Cotsia I, Zafeiriou SP. **ArcFace: Additive Angular Margin Loss for Deep Face Recognition**. IEEE Trans PAMI. 2021 Jun 9; doi: 10.1109/TPAMI.2021.3087709. https://github.com/deepinsight/insightface

Loss functions

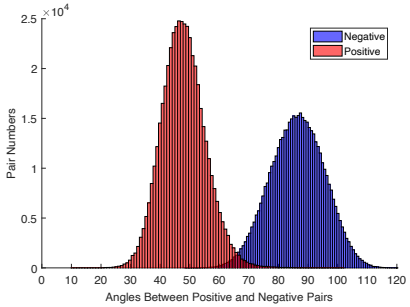
Loss Functions	LFW	CFP-FP	AgeDB-30
ArcFace (0.4)	99.53	95.41	94.98
ArcFace (0.45)	99.46	95.47	94.93
ArcFace (0.5)	**99.53**	**95.56**	**95.15**
ArcFace (0.55)	99.41	95.32	95.05
SphereFace [18]	99.42	-	-
SphereFace (1.35)	99.11	94.38	91.70
CosFace [37]	99.33	-	-
CosFace (0.35)	99.51	95.44	94.56
CM1 (1, 0.3, 0.2)	99.48	95.12	94.38
CM2 (0.9, 0.4, 0.15)	99.50	95.24	94.86
Softmax	99.08	94.39	92.33
Norm-Softmax (NS)	98.56	89.79	88.72
NS+Intra	98.75	93.81	90.92
NS+Inter	98.68	90.67	89.50
NS+Intra+Inter	98.73	94.00	91.41
Triplet (0.35)	98.98	91.90	89.98
ArcFace+Intra	99.45	95.37	94.73
ArcFace+Inter	99.43	95.25	94.55
ArcFace+Intra+Inter	99.43	95.42	95.10
ArcFace+Triplet	99.50	95.51	94.40

Table 2. Verification results (%) of different loss functions ([CASIA, ResNet50, loss*]).

Method	#Image	LFW	YTF
DeepID [32]	0.2M	99.47	93.20
Deep Face [33]	4.4M	97.35	91.4
VGG Face [24]	2.6M	98.95	97.30
FaceNet [29]	200M	99.63	95.10
Baidu [16]	1.3M	99.13	-
Center Loss [38]	0.7M	99.28	94.9
Range Loss [46]	5M	99.52	93.70
Marginal Loss [9]	3.8M	99.48	95.98
SphereFace [18]	0.5M	99.42	95.0
SphereFace+ [17]	0.5M	99.47	-
CosFace [37]	5M	99.73	97.6
MS1MV2, R100, ArcFace	5.8M	**99.83**	**98.02**



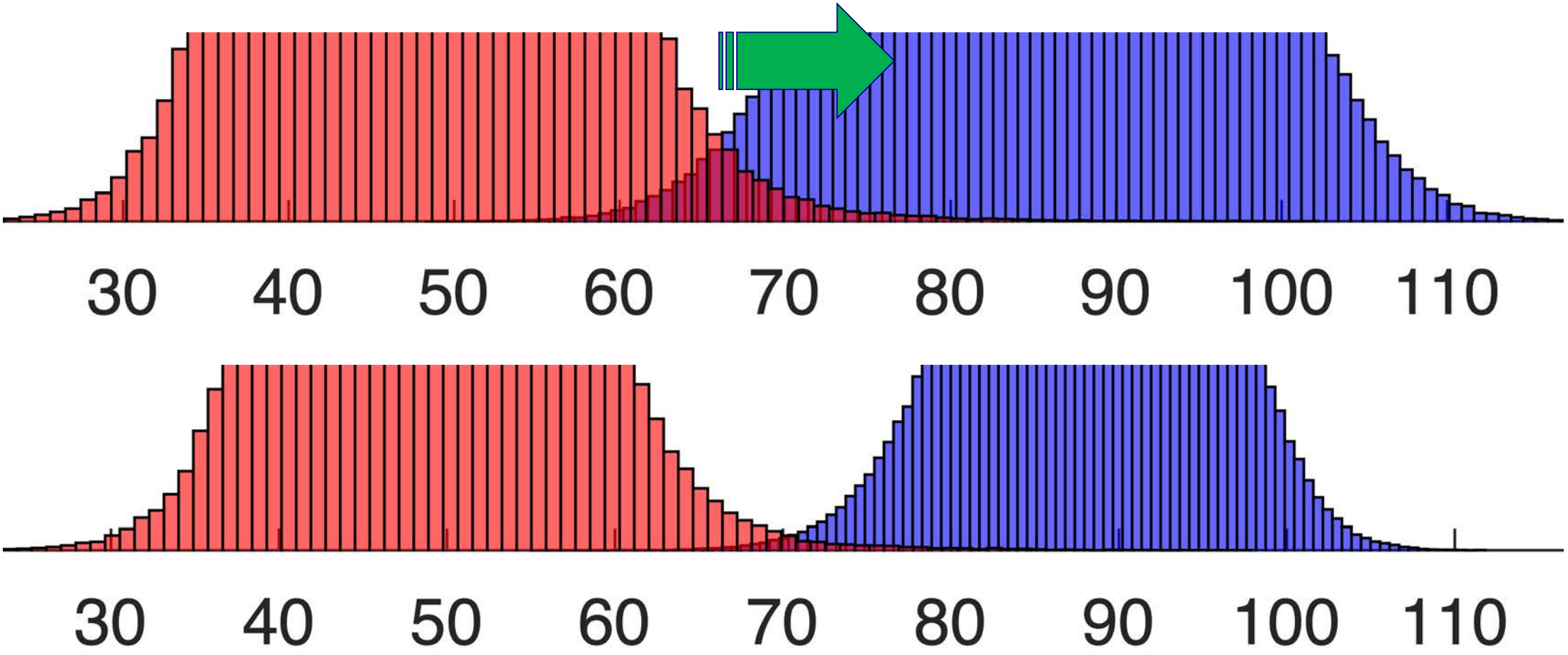
(a) ArcFace



(b) Triplet-Loss

Deng J, Guo J, Yang J, Xue N, Cotsia I, Zafeiriou SP. **ArcFace: Additive Angular Margin Loss for Deep Face Recognition**. IEEE Trans PAMI. 2021 Jun 9; doi: 10.1109/TPAMI.2021.3087709. https://github.com/deepinsight/insightface

Loss functions



Datasets

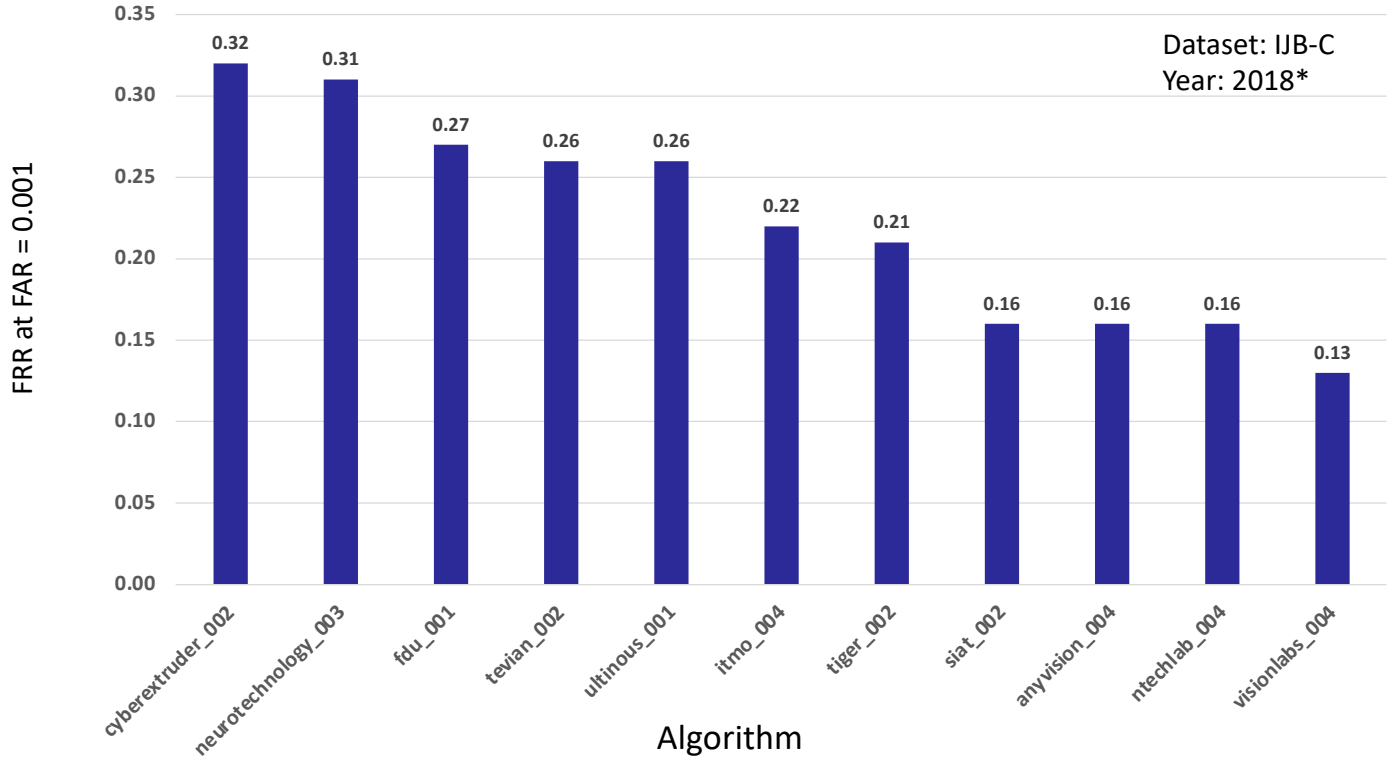
Dataset	Available	#Photos and #people
LFW	Public	13K of 5K people
CelebFaces 2014	Private	202K of 10K people
CASIA-WebFace 2014	Public	500K of 10K people
FaceScrub 2014	Public	100K of 500 people
YouTube Faces	Public	3425 videos of 1595 people
DeepFace (Facebook) 2014	Private	4.4 Million of 4K people
FaceNet (Google) 2015	Private	100-200 Million of 8M people
MegaFace	**Public**	**1 Million**

Figure 2: Representative sample of face recognition datasets that were created in the recent years (in addition to LFW). All the public datasets are small scale, and all the large scale datasets are mainly used for training rather than testing and are not publicly available. MegaFace (this paper) is the first large scale unconstrained dataset. It is collected from Flickr and will be available publicly.

Miller et al. (2015) Mega-Face: A million faces for recognition at scale.

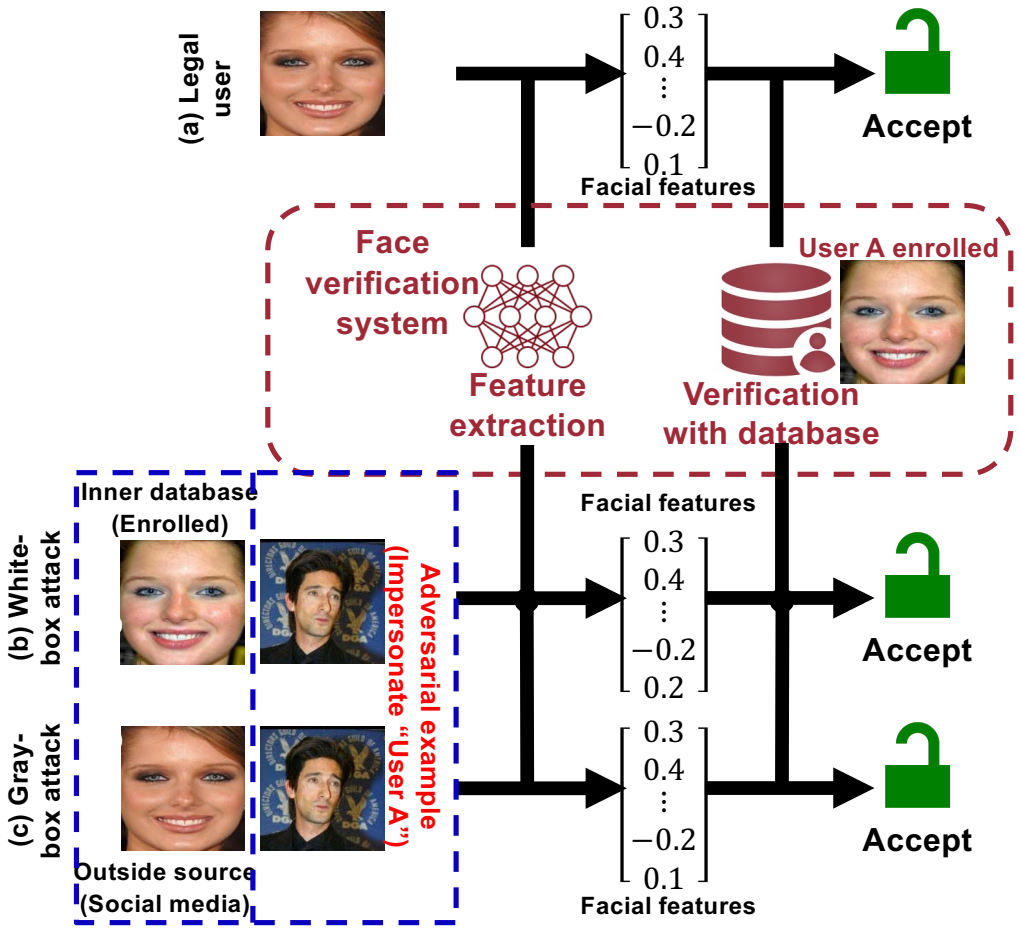
CNN Performance

FRVT 1:1 Wild-to-wild comparisons



Courtesy of J. Phillips (2021)

Adversarial Attacks

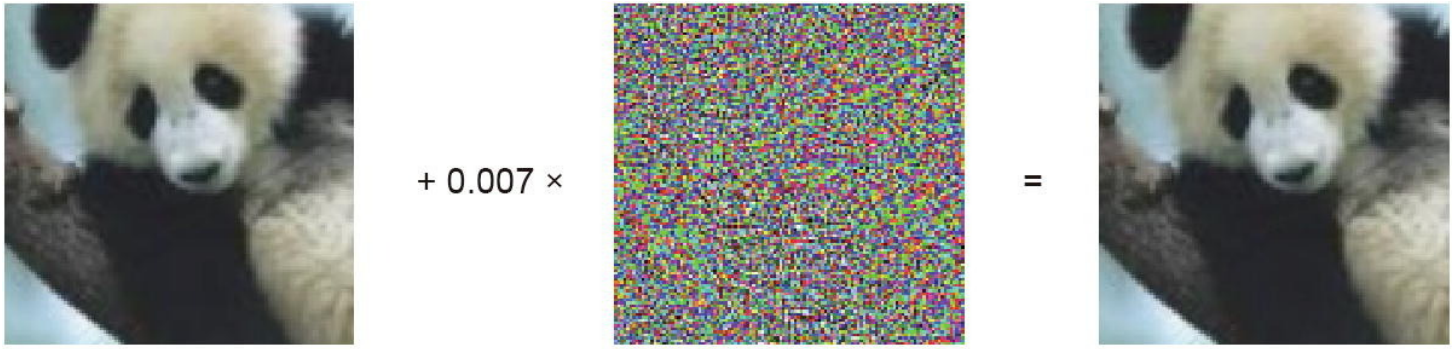


- **White-box attack**:
  - The deep-learning-based face verification model (feature extraction) is KNOWN to the attacker.
  - The precise enrolled face image is KNOWN to the attacker.
- **Gray-box attack**:
  - The feature extraction is KNOWN to the attacker.
  - The precise enrolled face image is UNKNOWN to the attacker.
- **Black-box attack**:
  - The deep-learning-based face verification model (feature extraction) is UNKNOWN to the attacker.

H. Wang, S. Wang, Z. Jin, Y. Wang, C. Chen, and M. Tistarelli, "**Similarity-based gray-box adversarial attack against deep face recognition**," in IEEE International Conference on Automatic Face and Gesture Recognition 2021 (FG2021), 2021

CNN Performance

- **A classic example**



Goodfellow IJ, Shlens J, Szegedy C. **Explaining and harnessing adversarial examples**. 6th International Conference on Learning Representations 2015. arXiv:1412.6572.

Kui Ren, Tianhang Zheng, Zhan Qin, Xue Liu, **Adversarial Attacks and Defenses in Deep Learning**, Engineering, Volume 6, Issue 3, 2020, Pages 346-360,

CNN Performance

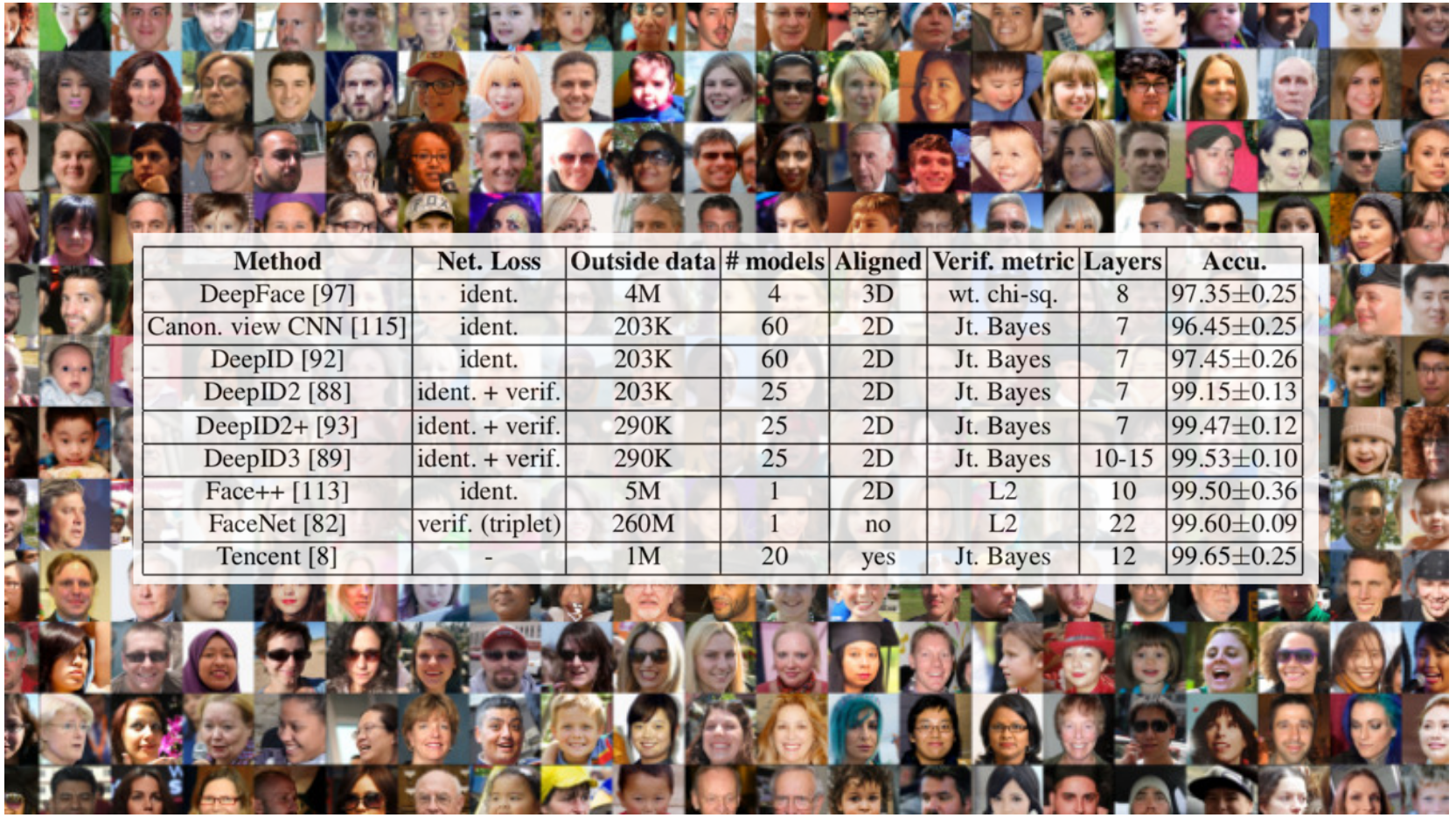
Deep Neural Network misclassifying stop sign to be speed limit 45 sign (left) using perturbations on stop sign

Eykholt K, Evtimov I, Fernandes E, Li B, Rahmati A, Xiao C, et al. **Robust physical-world attacks on deep learning visual classification**. In: Proceedings of the 2018 IEEE Conference on Computer Vision and Pattern Recognition; 2018 Jun 18–23; Salt Lake City, UT, USA; 2018. p. 1625–34.

Who is who?

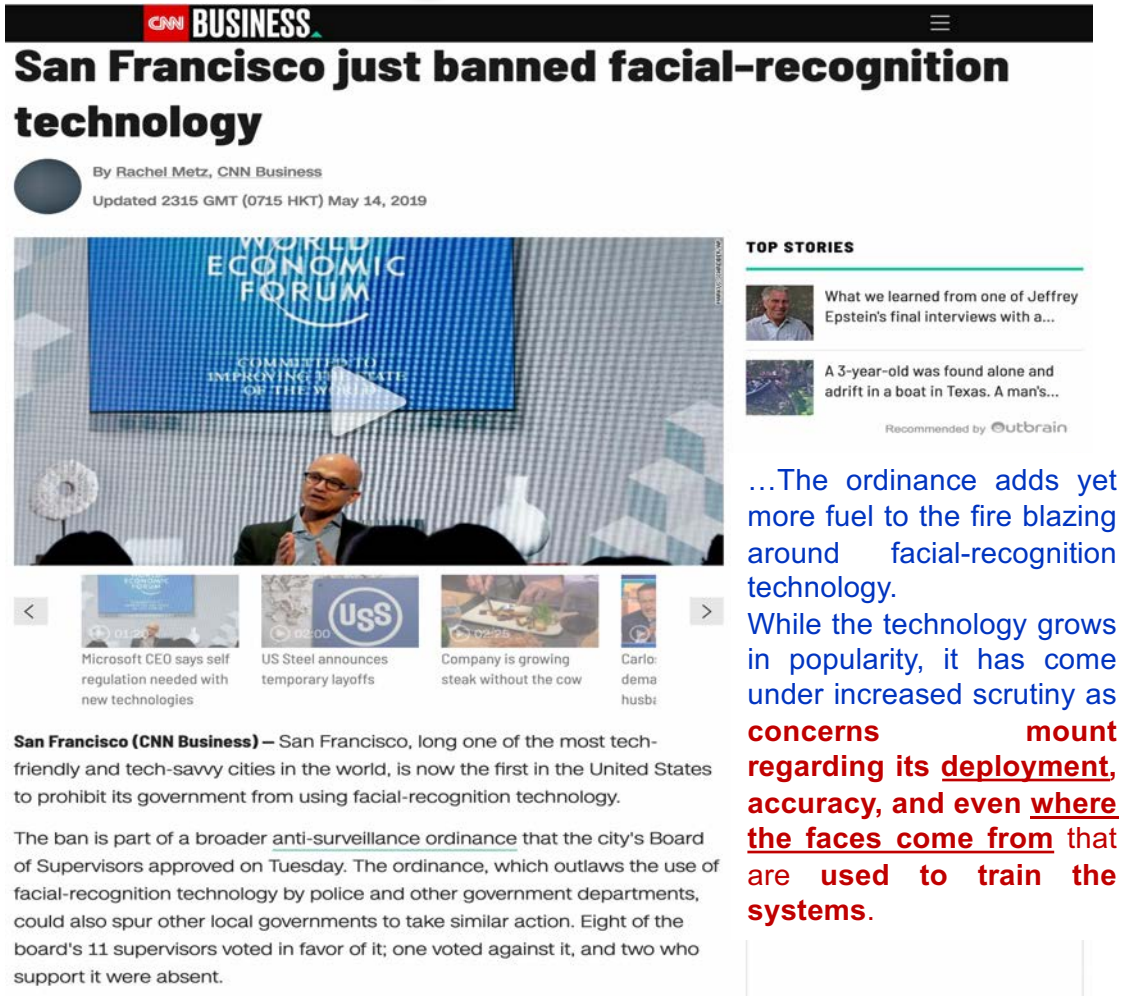
M. Sharif , S. Bhagavatula, L. Bauer, M. K. Reiter, "**Accessorize to a Crime: Real and Stealthy Attacks on State-of-the-Art Face Recognition**", CCS'16 October 24-28, 2016, Vienna, Austria

The "curse of training"



Method	Net. Loss	Outside data	# models	Aligned	Verif. metric	Layers	Accu.
DeepFace [97]	ident.	4M	4	3D	wt. chi-sq.	8	97.35±0.25
Canon. view CNN [115]	ident.	203K	60	2D	Jt. Bayes	7	96.45±0.25
DeepID [92]	ident.	203K	60	2D	Jt. Bayes	7	97.45±0.26
DeepID2 [88]	ident. + verif.	203K	25	2D	Jt. Bayes	7	99.15±0.13
DeepID2+ [93]	ident. + verif.	290K	25	2D	Jt. Bayes	7	99.47±0.12
DeepID3 [89]	ident. + verif.	290K	25	2D	Jt. Bayes	10-15	99.53±0.10
Face++ [113]	ident.	5M	1	2D	L2	10	99.50±0.36
FaceNet [82]	verif. (triplet)	260M	1	no	L2	22	99.60±0.09
Tencent [8]	-	1M	20	yes	Jt. Bayes	12	99.65±0.25

Face recognition concerns



CNN BUSINESS

San Francisco just banned facial-recognition technology

By Rachel Metz, CNN Business

Updated 2315 GMT (0715 HKT) May 14, 2019

Microsoft CEO says self regulation needed with new technologies

US Steel announces temporary layoffs

Company is growing steak without the cow

Carlo: dema husb

San Francisco (CNN Business) – San Francisco, long one of the most tech-friendly and tech-savvy cities in the world, is now the first in the United States to prohibit its government from using facial-recognition technology.

The ban is part of a broader anti-surveillance ordinance that the city's Board of Supervisors approved on Tuesday. The ordinance, which outlaws the use of facial-recognition technology by police and other government departments, could also spur other local governments to take similar action. Eight of the board's 11 supervisors voted in favor of it; one voted against it, and two who support it were absent.

…The ordinance adds yet more fuel to the fire blazing around facial-recognition technology.

While the technology grows in popularity, it has come under increased scrutiny as **concerns mount regarding its deployment, accuracy, and even where the faces come from** that are **used to train the systems**.

https://edition.cnn.com/2019/05/14/tech/san-francisco-facial-recognition-ban/index.html

CNNs: Where are we going?

International Journal of Computer Vision (2021) 129:781–802
https://doi.org/10.1007/s11263-020-01405-z

Deep Nets: What have They Ever Done for Vision?

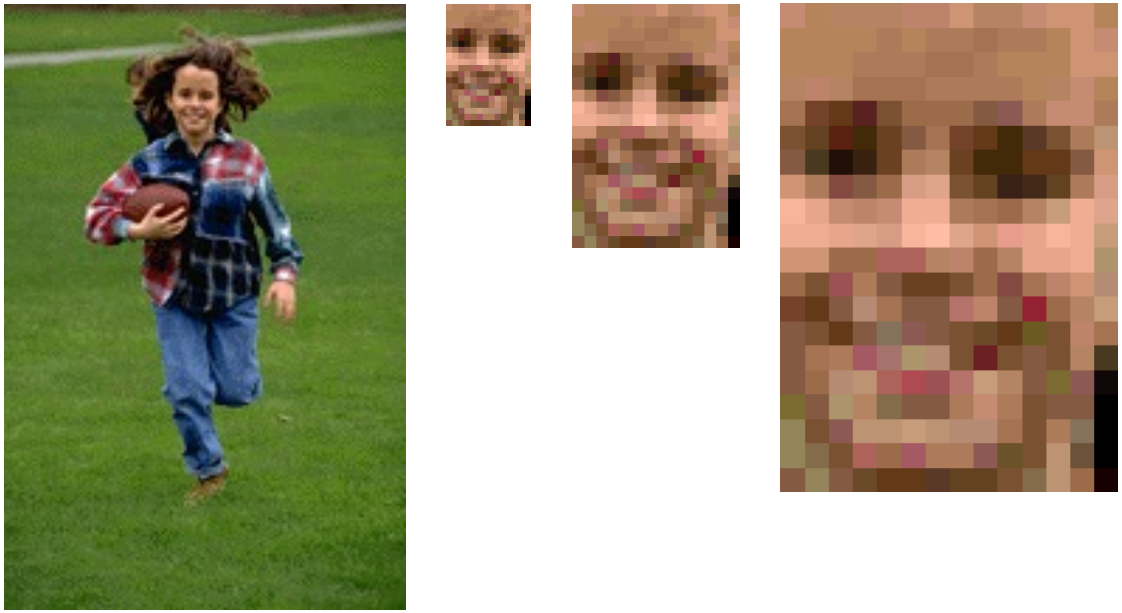
Alan L. Yuille[1] · Chenxi Liu[1]

Received: 10 January 2019 / Accepted: 9 November 2020 / Published online: 27 November 2020

Abstract
This is an opinion paper about the strengths and weaknesses of Deep Nets for vision. They are at the heart of the enormous recent progress in artificial intelligence and are of growing importance in cognitive science and neuroscience. They have had many successes but also have several limitations and there is limited understanding of their inner workings. At present Deep Nets perform very well on specific visual tasks with benchmark datasets but they are much less general purpose, flexible, and adaptive than the human visual system. We argue that Deep Nets in their current form are unlikely to be able to overcome the fundamental problem of computer vision, namely how to deal with the combinatorial explosion, caused by the enormous complexity of natural images, and obtain the rich understanding of visual scenes that the human visual achieves. We argue that **this combinatorial explosion takes us into a regime where "big data is not enough" and where we need to rethink our methods for benchmarking performance and evaluating vision algorithms**. We stress that, as vision algorithms are increasingly used in real world applications, that performance evaluation is not merely an academic exercise but has important consequences in the real world. It is impractical to review the entire Deep Net literature so we restrict ourselves to a limited range of topics and references which are intended as entry points into the literature. The views expressed in this paper are our own and do not necessarily represent those of anybody else in the computer vision community.

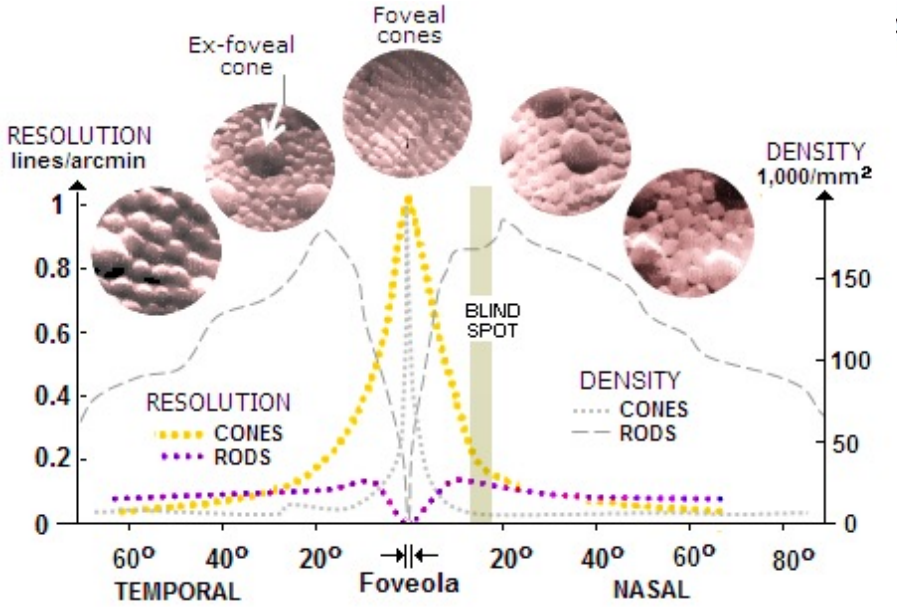
Keywords Deep neural networks · Computer vision · Success · Limitation · Cognitive science · Neuroscience

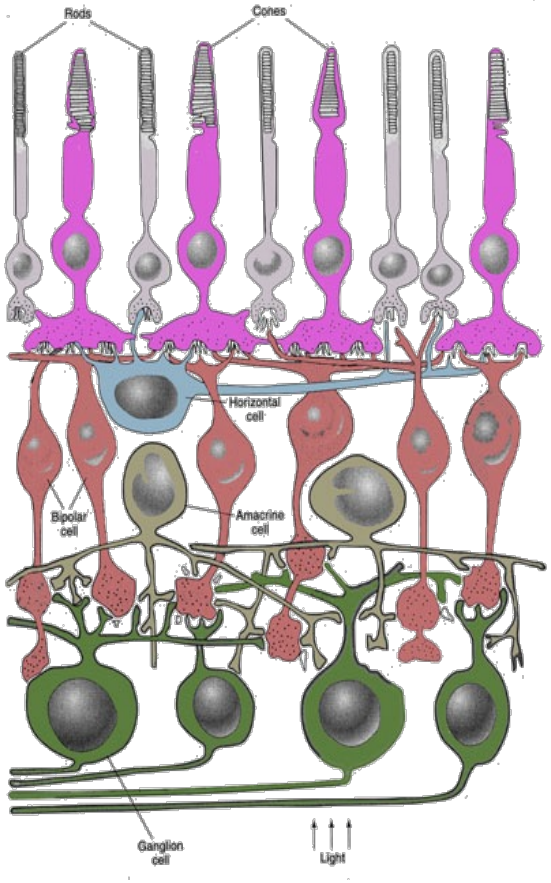
A different "perspective"



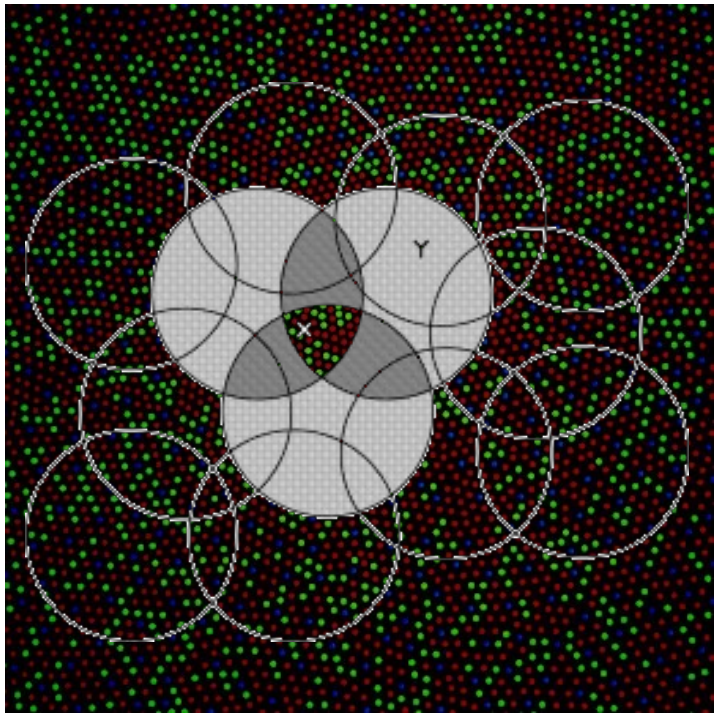
Spatial distribution and **Frequency tuning**

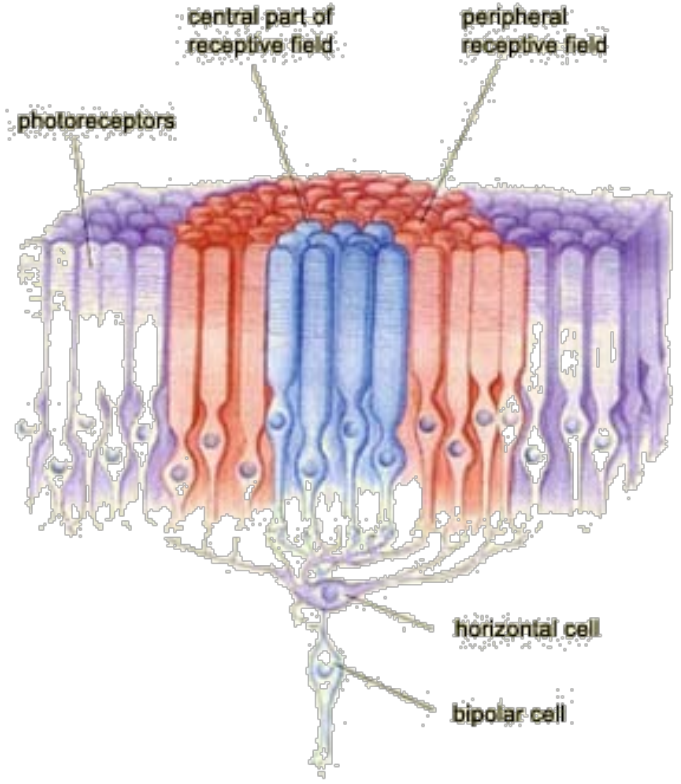
The human retina





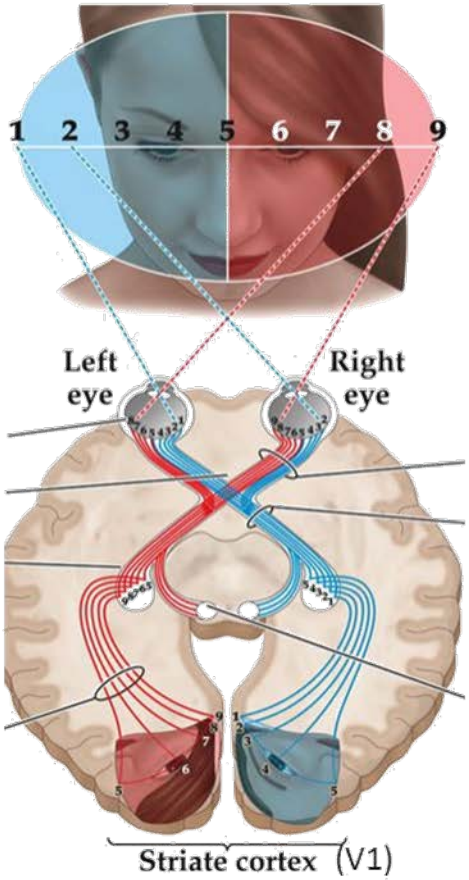
Receptive fields

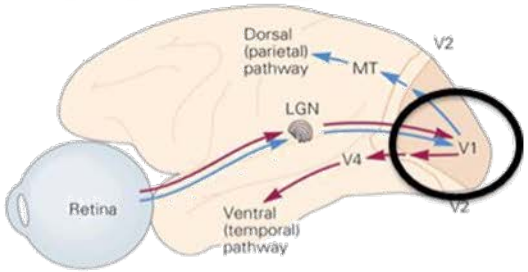




Retinotopic mapping

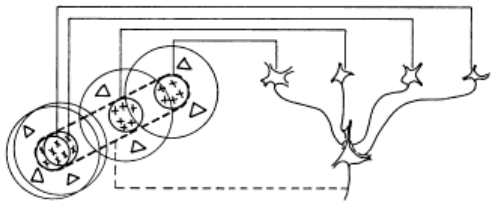
V1 retinotopic maps



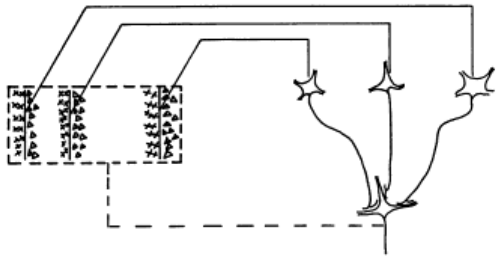


- Each point of the visual field maps on to a local group of neurons in V1.
- Retinotopy = Remapping of retinal image onto cortical surface
- Foveal region uses more of V1 (greater magnification factor)

Hubel & Wiesel 1962



Text-fig. 19. Possible scheme for explaining the organization of simple receptive fields. A large number of lateral geniculate cells, of which four are illustrated in the upper right in the figure, have receptive fields with 'on' centres arranged along a straight line on the retina. All of these project upon a single cortical cell, and the synapses are supposed to be excitatory. The receptive field of the cortical cell will then have an elongated 'on' centre indicated by the interrupted lines in the receptive-field diagram to the left of the figure.

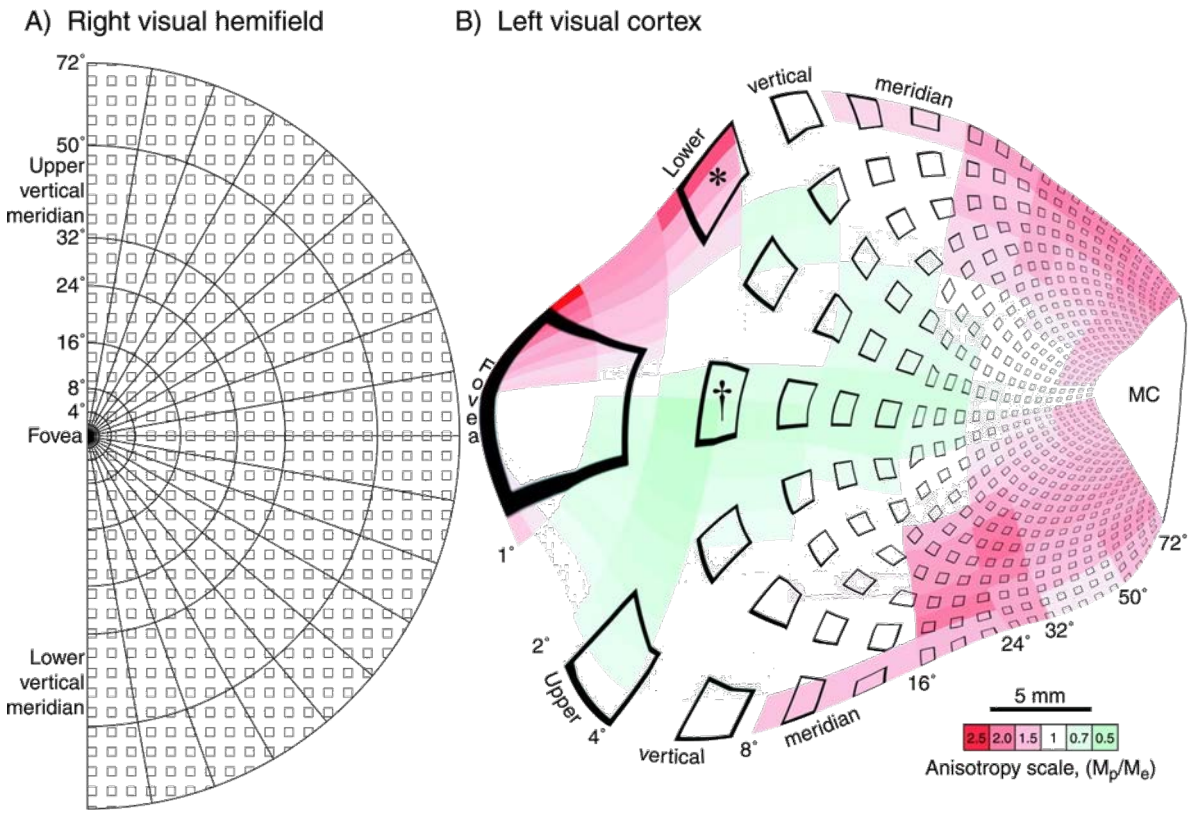


Text-fig. 20. Possible scheme for explaining the organization of complex receptive fields. A number of cells with simple fields, of which three are shown schematically, are imagined to project to a single cortical cell of higher order. Each projecting neurone has a receptive field arranged as shown to the left: an excitatory region to the left and an inhibitory region to the right of a vertical straight-line boundary. The boundaries of the fields are staggered within an area outlined by the interrupted lines. Any vertical-edge stimulus falling across this rectangle, regardless of its position, will excite some simple-field cells, leading to excitation of the higher-order cell.

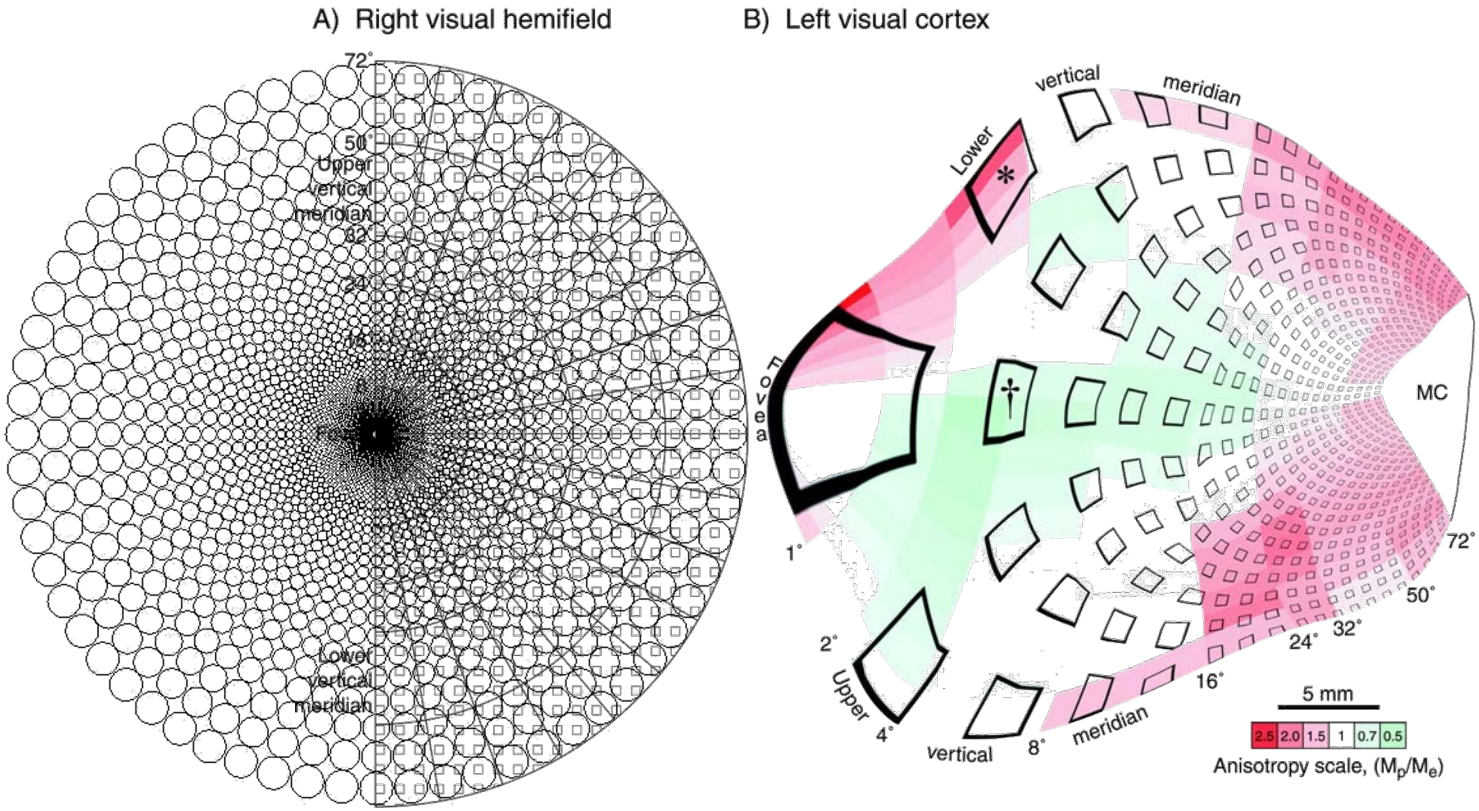
Simple and Complex cells

Hubel DH & Wiesel TN (1962). "**Receptive fields, binocular interaction and functional architecture in the cat's visualcortex**". JPhysiol160, 106–154

Retinotopic mapping

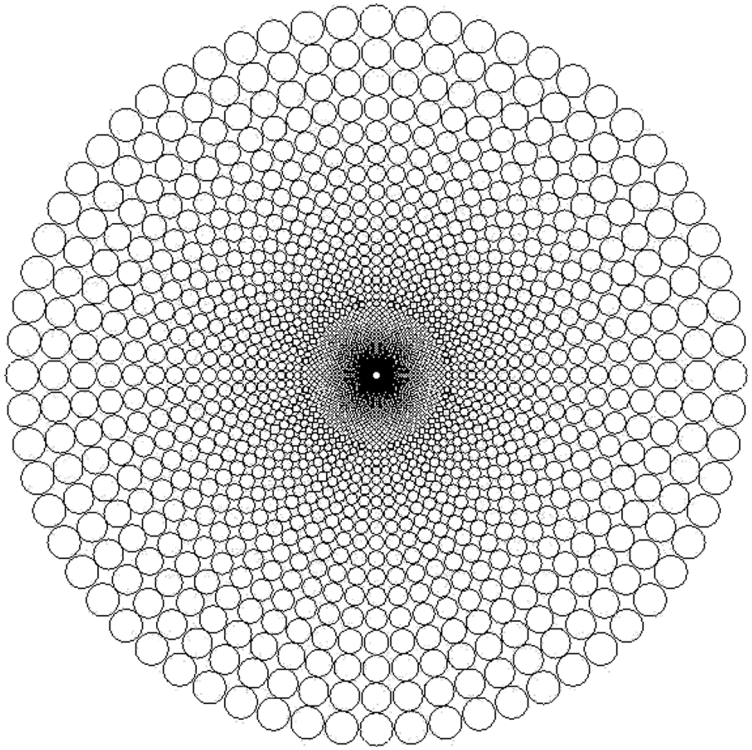


Retinotopic mapping



Log-Polar mapping

The **complex log-polar transform** is a good approximation of the retinal sampling



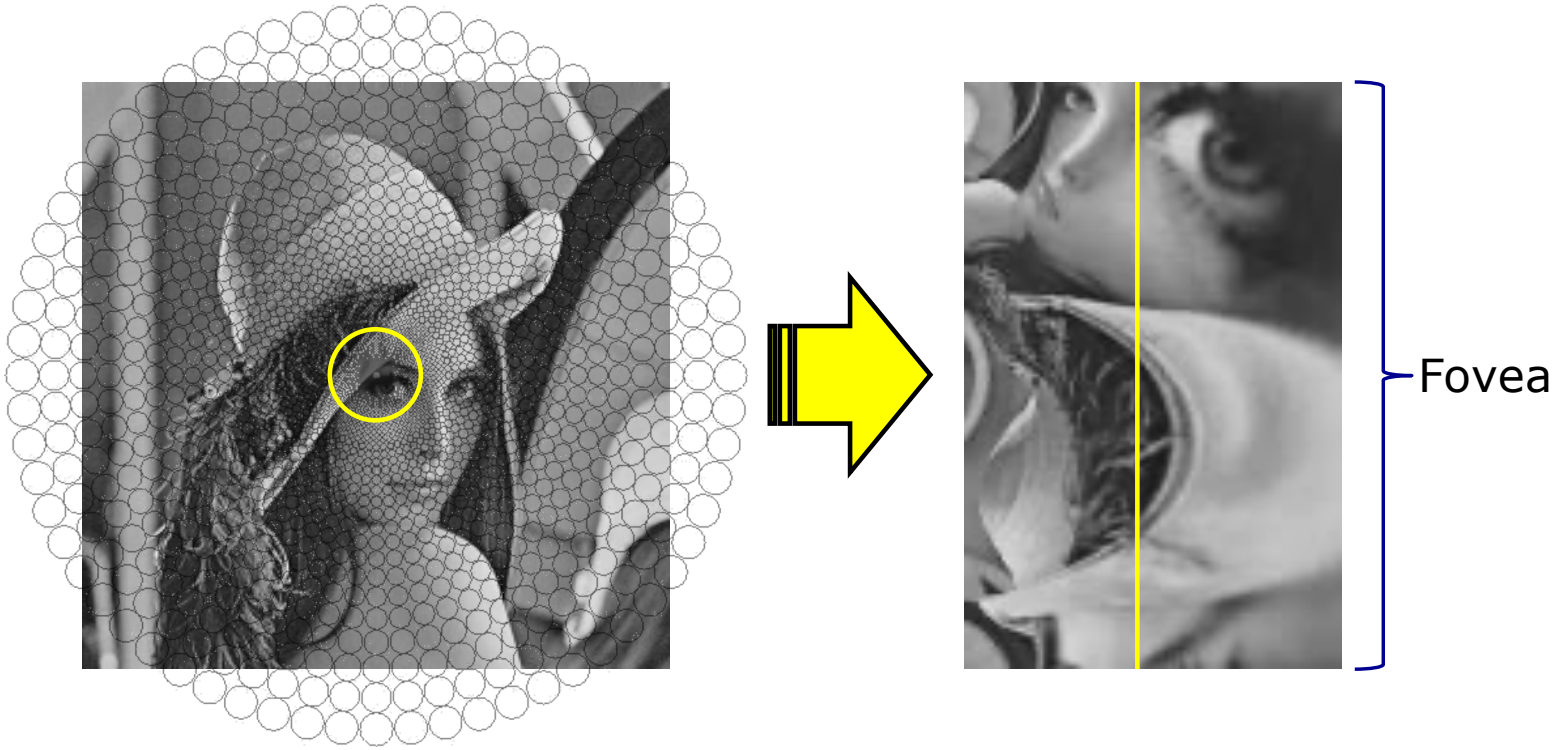
$$\begin{cases} x = \rho \sin\theta \\ y = \rho \cos\theta \end{cases}$$

$$\begin{cases} \xi = \log_a\left(\rho / \rho_0\right) \\ \eta = q\theta \end{cases}$$

Massone, L., Sandini,G. and Tagliasco, V. "**Form-invariant topological mapping strategy for 2-d shape recognition**", CVGIP, vol. 30 No.2, pp. 169-188, 1985

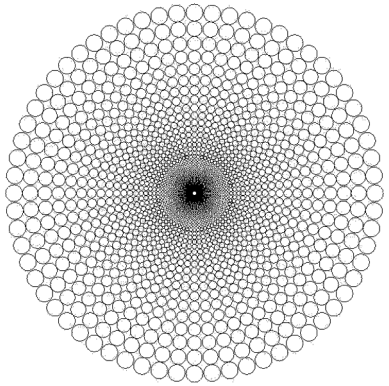
Log-Polar mapping

The **complex log-polar transform** is a good approximation of the retinal sampling



Massone, L., Sandini,G. and Tagliasco, V. "**Form-invariant topological mapping strategy for 2-d shape recognition**", CVGIP, vol. 30 No.2, pp. 169-188, 1985

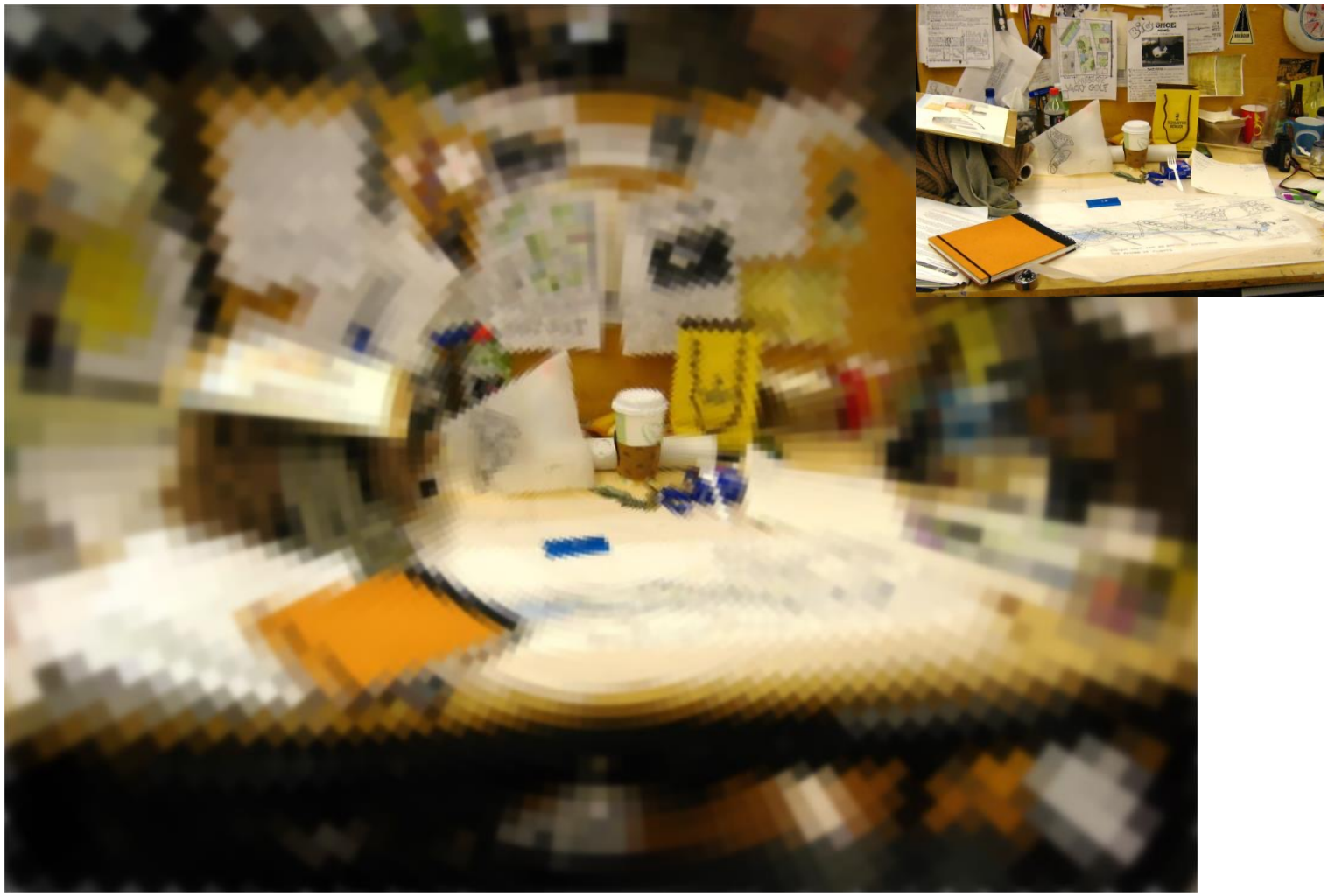
Space-variant imaging



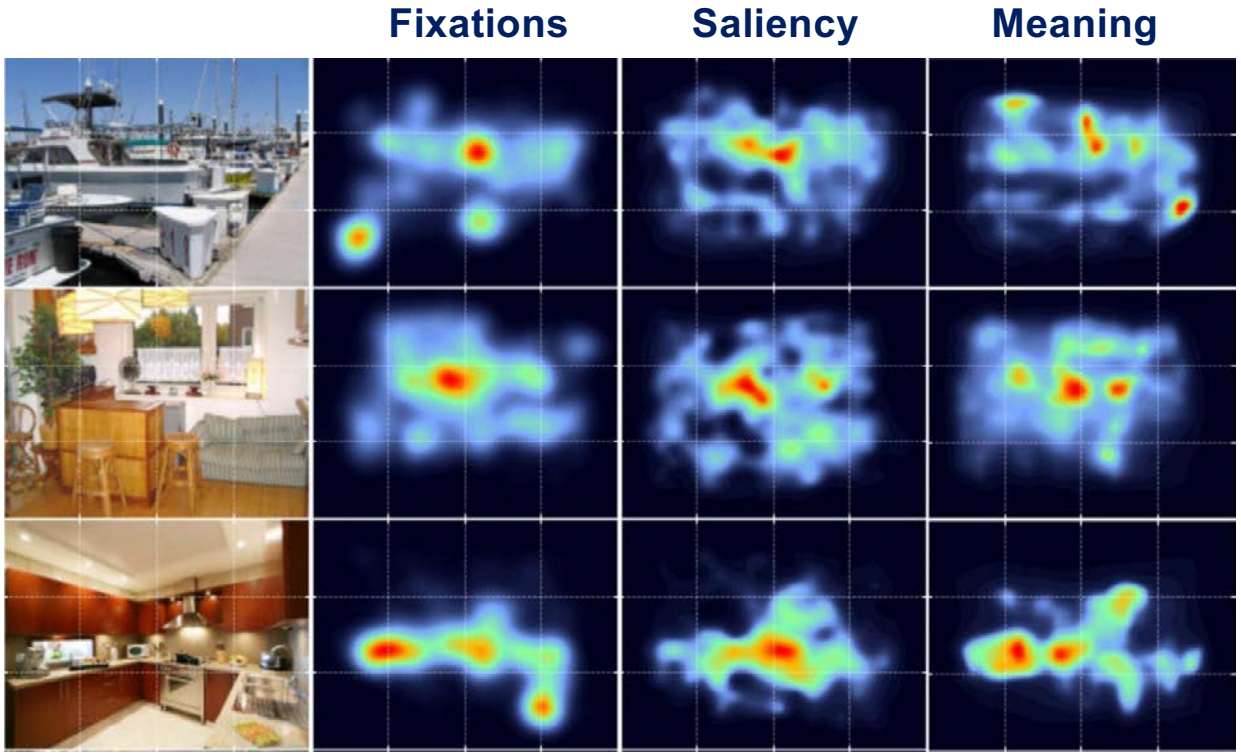
Tistarelli, M. and Grosso, E. (1997) "**Active face recognition with an hybrid approach**" *Pattern Recognition Letters*, Vol. 18, pp 933-946, 1997

Tistarelli, M. and Grosso, E. (2000) "**Active vision-based face authentication**" *Image and Vision Computing*, Vol. 18, no. 4, pp 299-314, 2000

Visual attention



Visual attention

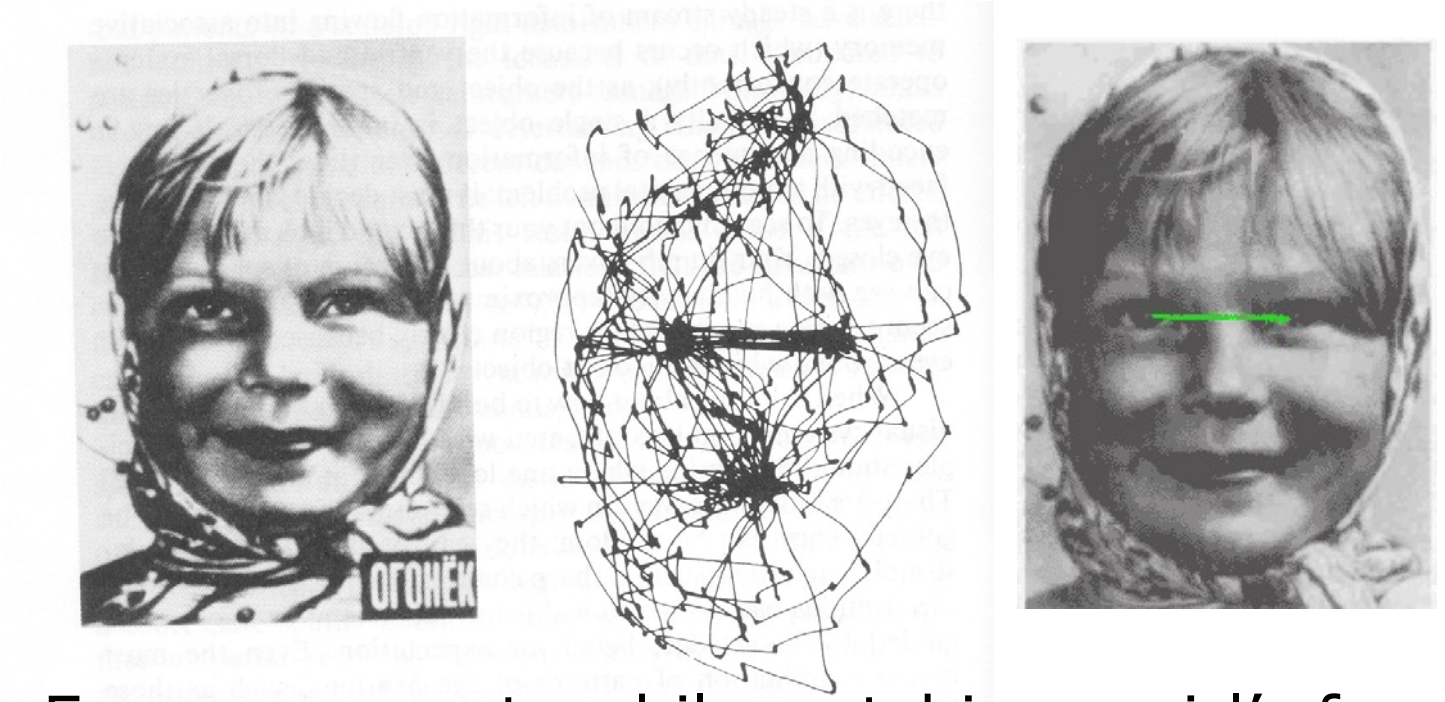


- Attention is driven by **utilitarian features** related to the **objects' meaning**

J.M,. Henderson, T.R. Hayes, **Meaning guides attention in real-world scene images: Evidence from eye movements and meaning maps**, Journal of Vision 18(6):1-18, June 2018

Just look once...

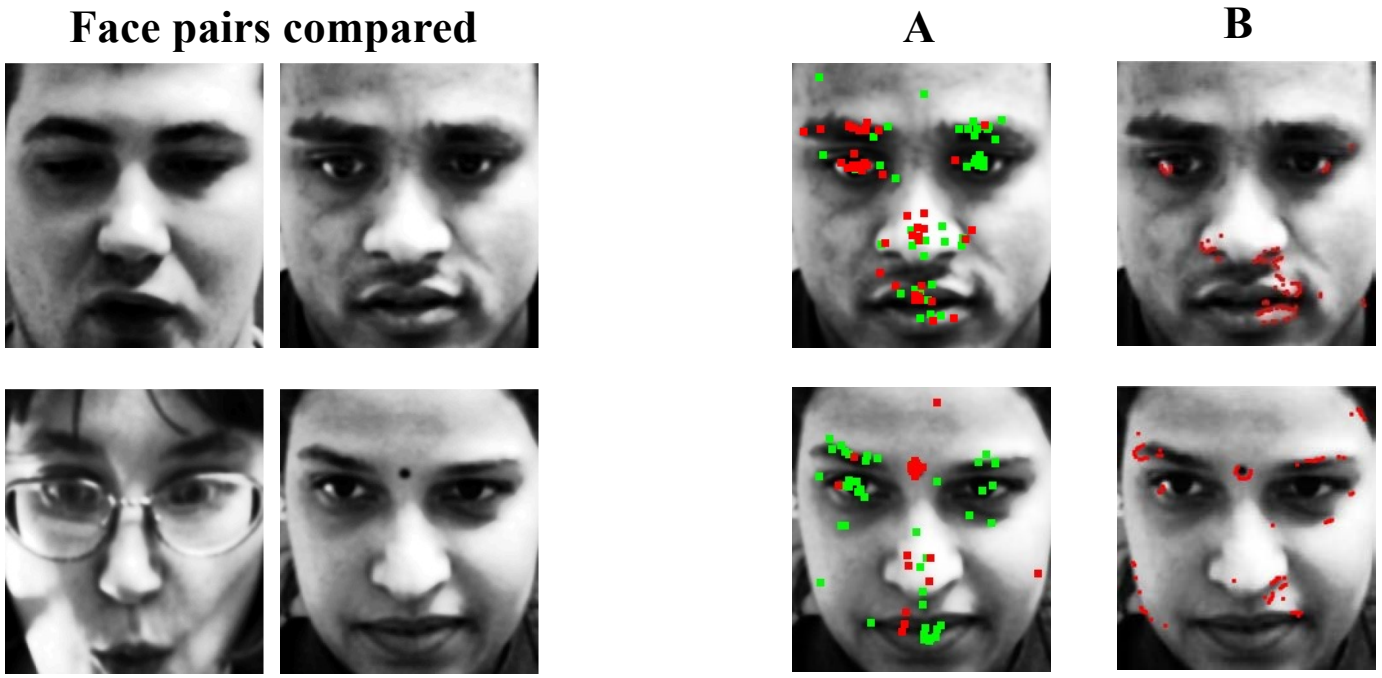
Visual attention



Eye movements while watching a girl's face

A.L. Yarbus, "**Eye Movements and Vision**", Plenum Press, 1967

Visual attention in face comparison

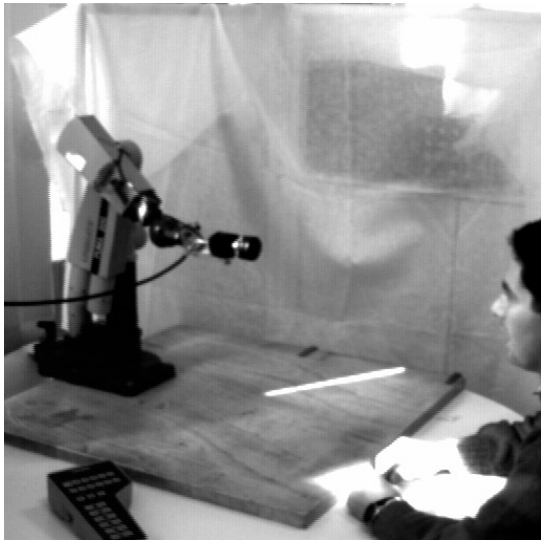


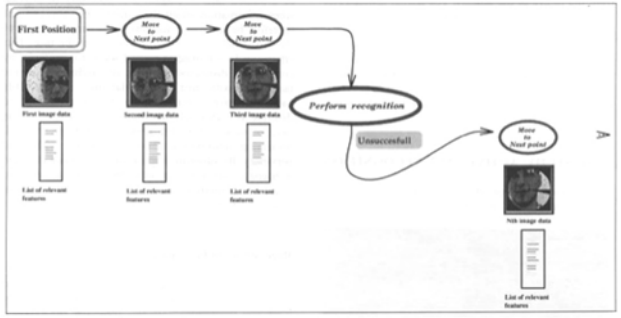
(A) perceptual and (B) computational results of saliency of local facial features, demonstrate the relevance of *non-standard* facial landmarks

Bicego M., Brelstaff G., Brodo L., Grosso E., Lagorio A. and Tistarelli M. (2007) "**Distinctiveness of faces: a computational approach**", ACM Transactions on Applied Perception, Vol. 5, n. 2, 2008.

Once upon a time…

June 1994… Seattle, WA





Recognition by Using an Active/Space-Variant Sensor

M. Tistarelli
DIST - University of Genoa
Laboratory for Integrated Advanced Robotics (LIRA - Lab)
via Opera Pia 11a - 16145 Genoa, Italy

Abstract

The problem of object recognition is addressed. In the literature this task has been generally considered in a "passive" perspective, where everything is static and there is no definite relation between the object and its environment. We propose an "active" approach for object recognition, based on the capability of the observer to move and give a better description of the object under consideration and also to take advantage of the relations between the objects and the environment. This can be accomplished at the task level and at the sensor level.

The face recognition problem, based on the face-space approach, is considered to demonstrate the advantage of adopting an active retina to sample the face, build a database and perform the recognition task. By using an active space-variant retina the size of the database is considerably reduced and consequently also the processing time for recognition.

A comparative experiment using the active and static approach is presented.

1 Introduction

Object recognition is one of the most "classical" themes in artificial intelligence applied to vision. Nonetheless up to now it is not a solved problem at all, but many different systems and methods have been investigated with limited success [1]. Certainly the reason of such effort is the formidable complexity of the recognition problem and the ability of humans to recognize objects quite quickly. But, is this ability due to a particular efficency of the search strategy in the model database? or is it due to the computational power of the inference engine (the brain)? without any doubt these are two relevant characteristics of the human brain, but these are not necessarily the primary reasons for the efficency of the human visual system. On the other hand we can consider that all the research carried out in the past, along these directions, did not obtain the expected results.

[1] The success of the techniques is limited in the sense that the generality of the solutions is not even comparable to the aims, which is to develop a fully general object recognition system, working in real world environments.

2 Fixation and recognition

What is the role of fixation in the recognition process? Yarbus, in his work on ocular movements [8], demonstrated that the sequence of fixations performed by the human oculo motor system, strongly depends on the task (in this case the question asked to the subject). He also showed that the eyes perform a particular sequence of fixations, if the subject has to recognize a part or a person in the scene. The eyes are successively directed toward the parts of the scene containing the most relevant features [2]. This motion strategy suggests that the motion of the eyes is particularly important for recognition (at least in the human visual system).

It is generally assumed that, for recognition, it is desirable to have a high resolution description of the most salient features of the interest object. This can be accomplished either by "foveating", in rapid succession, these parts of the scene or moving an interest window on a high resolution image [10].

Certainly object features are important for recognition, but the context, or the peripheral part of the visual field, allows to define a spatial relation among the object features which really characterize the object itself. A way to meet these requirements is to adopt a space-variant sampling strategy of the image, where the central part of the visual field is sampled at a higher resolution than the periphery, with a linear variation in resolution from the center to the periphery. An advantage of this approach is the great data

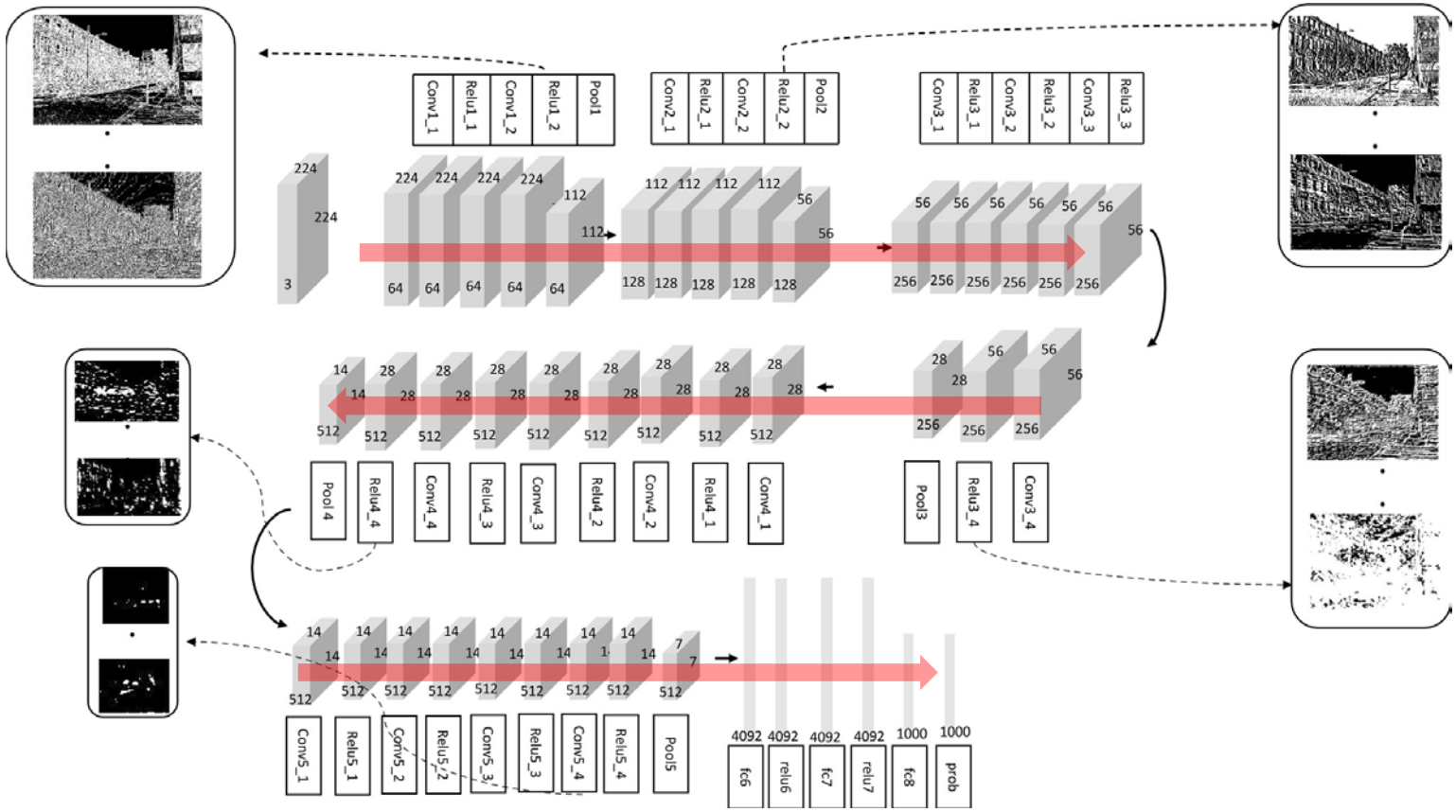
[2] Eklundh [9] demonstrated that these points can be recovered through a scale-space analysis of the image.

Tistarelli, M (1994) "**Recognition by using an active/space-variant sensor**" *IEEE CVPR*, 1994

Tistarelli, M. and Grosso, E. (1997) "**Active face recognition with an hybrid approach**" *Pattern Recognition Letters*, Vol. 18, pp 933-946, 1997

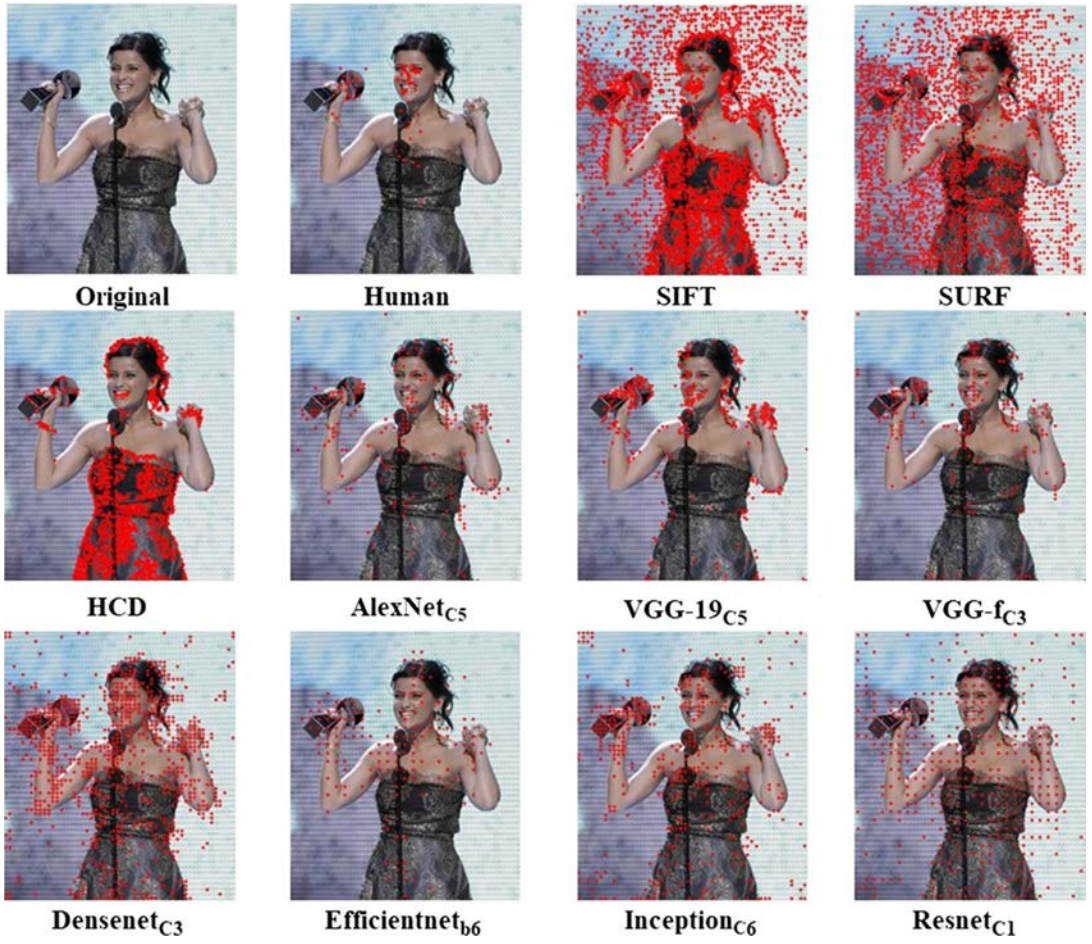
Tistarelli, M. and Grosso, E. (2000) "**Active vision-based face authentication**" *Image and Vision Computing*, Vol. 18, no. 4, pp 299-314, 2000

Do CNNs exploit Visual Attention?



M. Cadoni, A. Lagorio, S. Khellat-Kihel, E. Grosso (2021) "**On the correlation between human fixations, handcrafted and CNN features**", Neural Computing and Applications. https://doi.org/10.1007/s00521-021-05863-5.

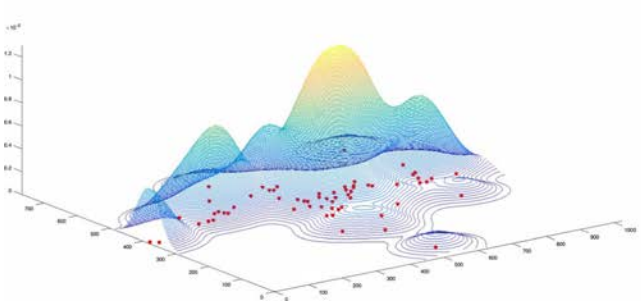
CNNs vs Human Visual Attention

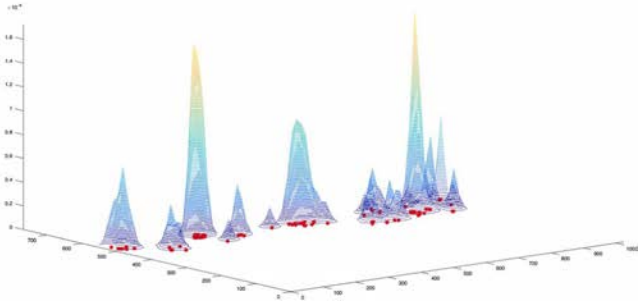


CNNs vs Human Visual Attention

Fixation points

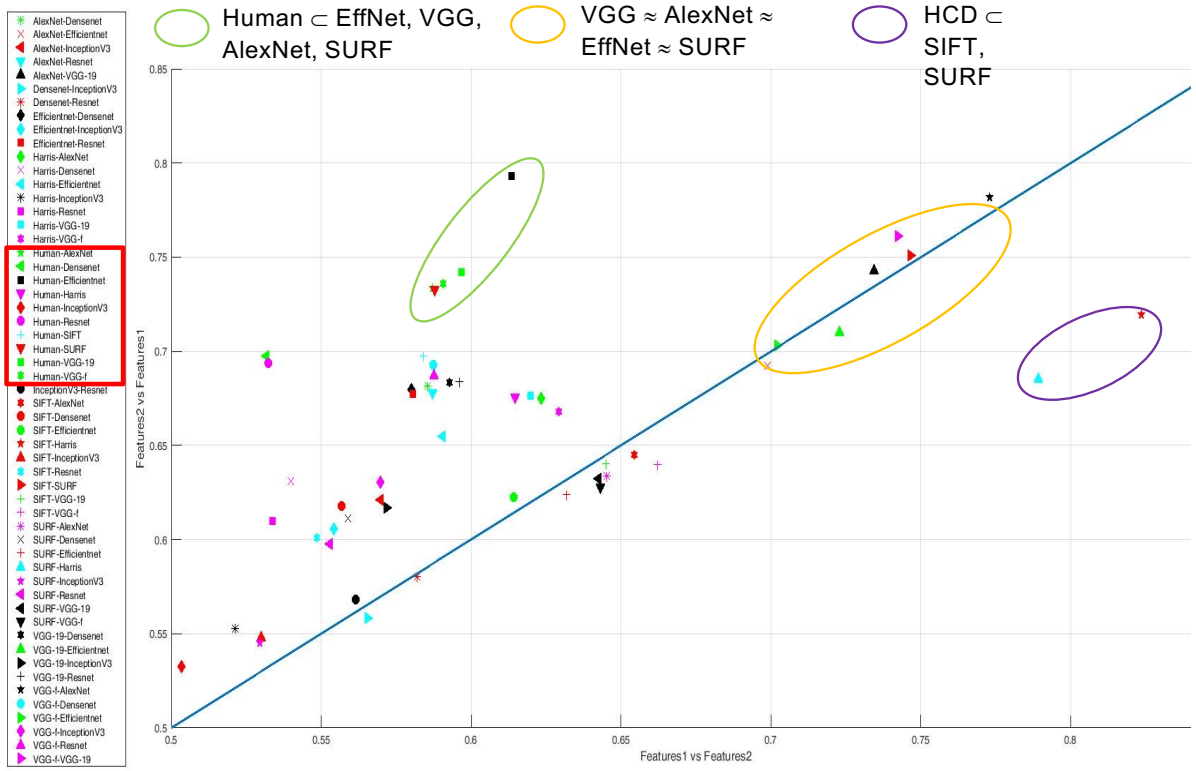
AlexNet interest points.





Interest regions are modeled via **K**erne**l** **D**ensity **E**stimation.

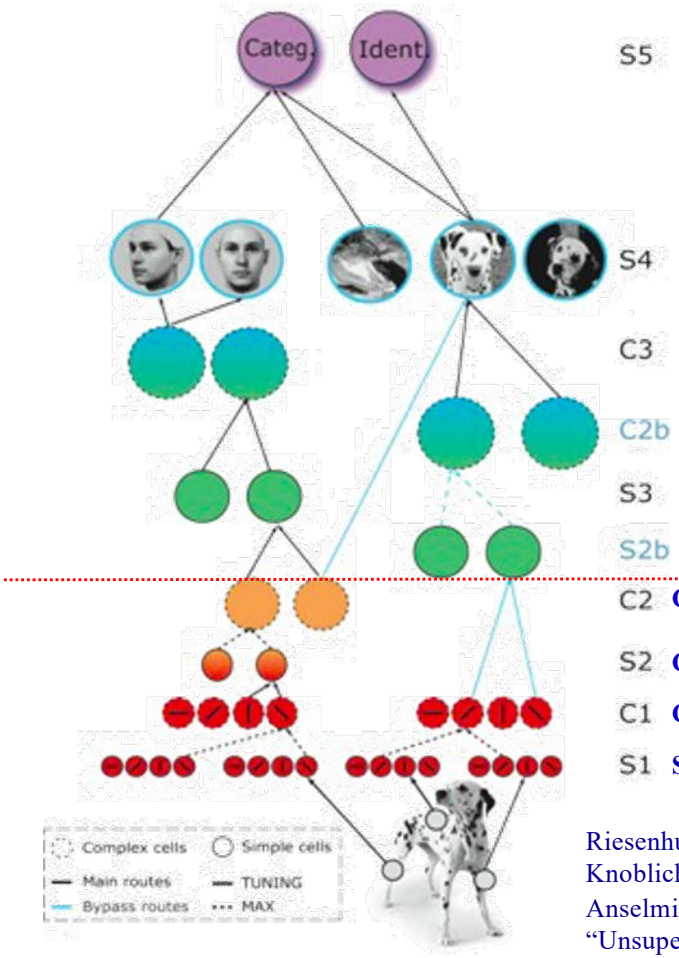
CNNs vs Human Visual Attention

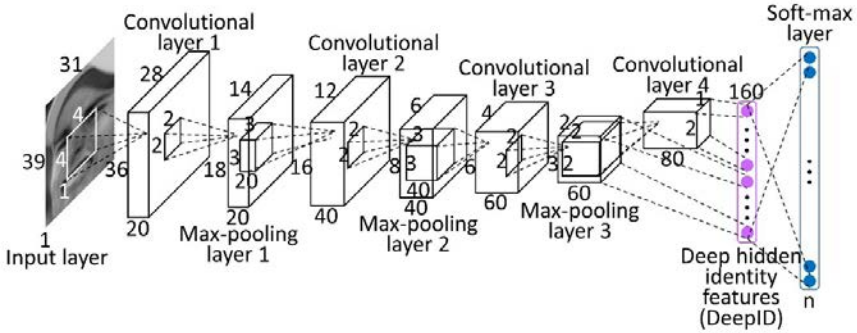


Local similarity between human fixations, CNNs and handcrafted features

M. Cadoni, A. Lagorio, E. Grosso, T. Jia Huei, C. Chee Seng (2021) "**From early biological models to CNNs: do they look where humans look?**", 25th Int.l Conference on Pattern Recognition ICPR 2020, pp. 6313-6320. doi: 10.1109/ICPR48806.2021.9412717.

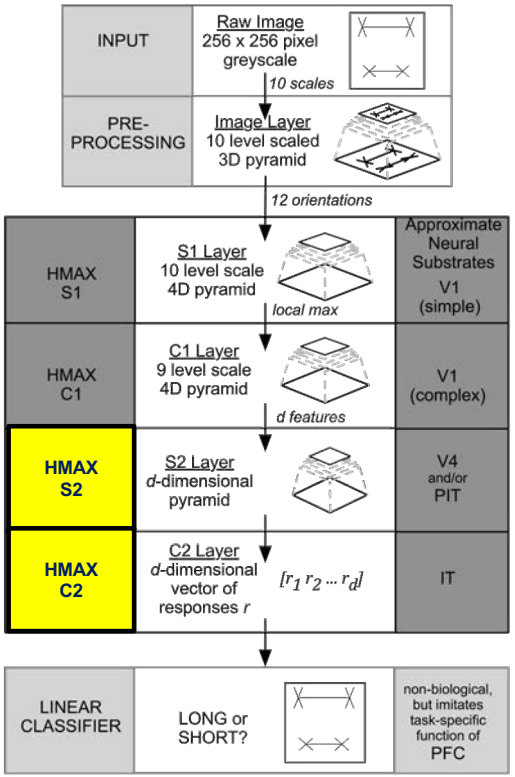
Brain models

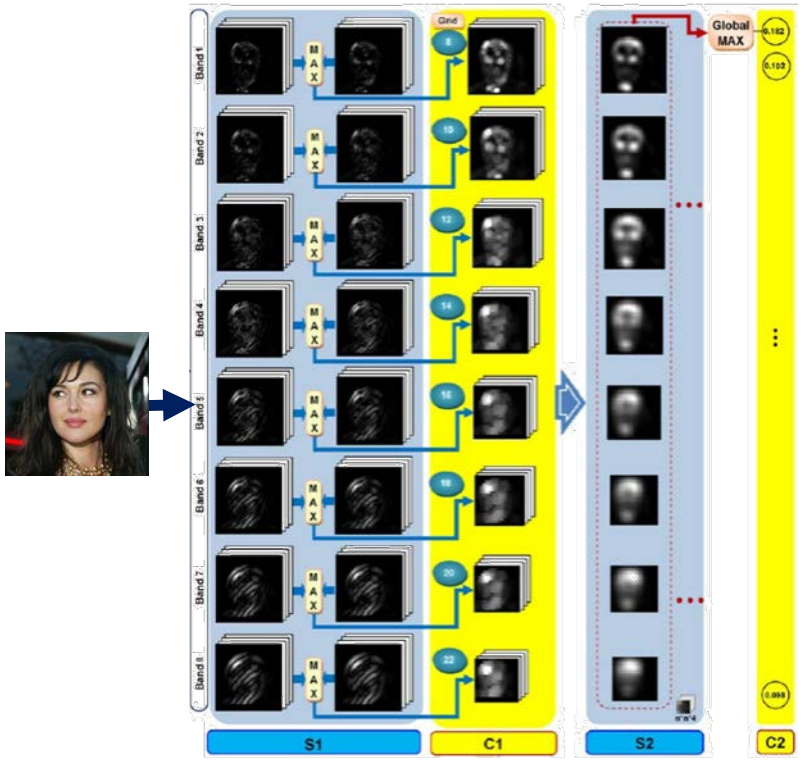




Riesenhuber & Poggio 1999, 2000; Serre Kouh Cadieu
Knoblich Kreiman & Poggio 2005; Serre Oliva Poggio 2007
Anselmi, F., Leibo, J. Z., Rosasco, L., Mutch, J., Tacchetti, A., and Poggio, T.,
"Unsupervised learning of invariant representations", Theoretical Computer Science, 2015.

The HMAX model

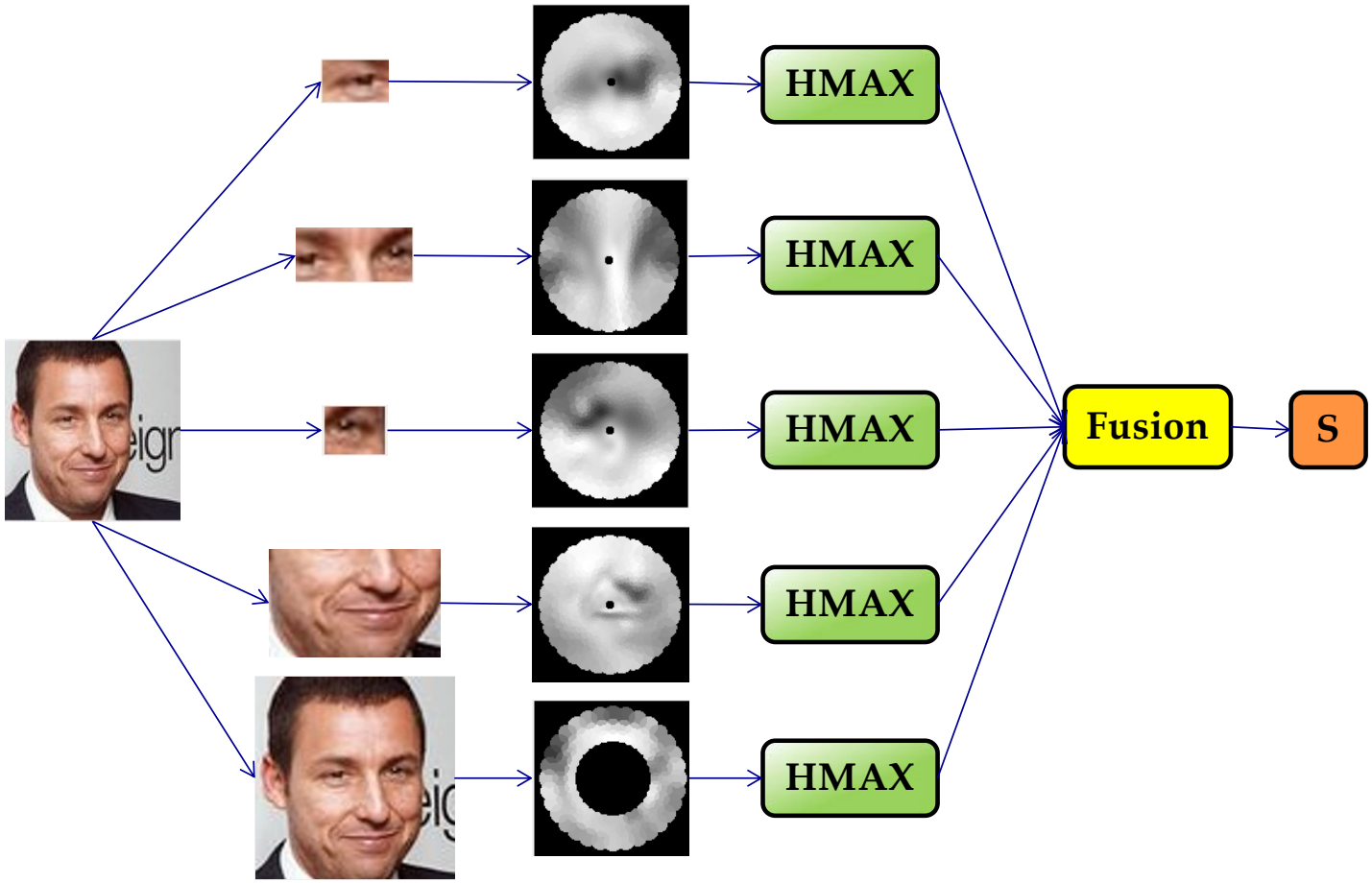




Riesenhuber, M. & Poggio, T. (1999). Hierarchical Models of Object Recognition in Cortex. Nature Neuroscience 2: 1019-1025.

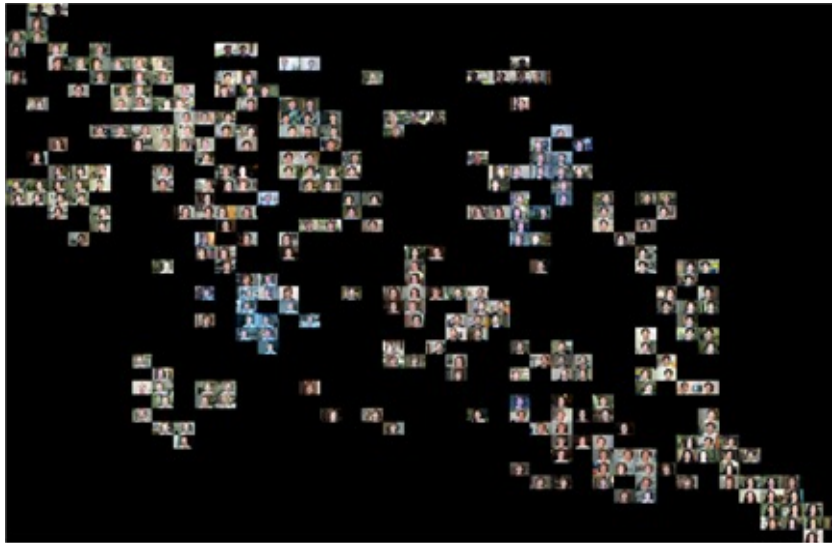
(S1) In this layer an input image is analyzed with a pyramid of filters (16 filter sizes×4 orientations = 64 images)
(C1) In this layer, the local maximum between 2 adjacent scales with the same orientation is taken.
(S2) The Euclidean distances between stored prototypes, which are obtained in the learning stage, and new input is computed.
This process occurs for all bands in C1 and as a result, S2 maps are obtained.
(C2) The global maximum is computed over all S2 responses in all positions and scales in this layer.

Foveated HMAX



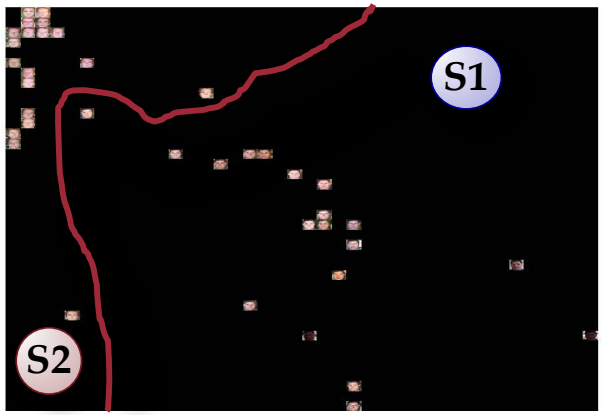
Foveated face recognition

HMAX Space representation on uniformly sampled face images



HMAX Space representation on log-polar sampled face images

Foveated face recognition



Uniform resolution

Log-polar mapping

Training	Testing	FF	SRC	MSSRC	VGG	Outer face	Ocular regions	Fusion
1 *Lab light*	2 *Dim light*	54.48	52.79	47.21	**62.27**	53.15	33.33	54.95
1 *Lab light*	3 *Sun light*	45.27	51.18	46.15	49.09	94.31	91.87	**95.12**
2 *Dim light*	1 *Lab light*	25.52	44.18	43.06	50.91	56.76	66.67	**78.38**
2 *Dim light*	3 *Sun light*	56.80	58.58	60.36	38.18	84.68	73.87	**84.68**
3 *Sun light*	1 *Lab light*	24.77	17.64	17.64	47.27	48.78	73.17	**73.98**
3 *Sun light*	2 *Dim light*	**56.01**	51.95	45.85	33.64	48.65	31.53	50.45

Performances are compared with Fisher Faces (FF), Sparse Representation based Classification (SRC), Mean-Sequence SRC (MSSRC) and VGG deep CNN.

S. Khellat Khiel, A. Lagorio, M. Tistarelli. "**Face Recognition 'On the Move' Combining Incomplete Information**". Proc. of 6th Int.l Workshop on Biometrics and Forensics, June 7,8 2018, Alghero, Italy. IEEE 2018.

S. Khellat Khiel, A. Lagorio, M. Tistarelli. "**Foveated vision for biologically-inspired continuous face authentication**". In A. Rattani Ed. *Selfie Biometrics: Methods and Challenges*, Springer 2019.

Vision Transformers

- State of the art Deep Learning architecture.
- Designed for Natural Language Processing and then extended to Vision [1].

  Applied to: face recognition [2], object recognition [3] and presentation attack detection [4] with SoTA results.
- **Key concept** Multi-Headed Self Attention (MHSA).

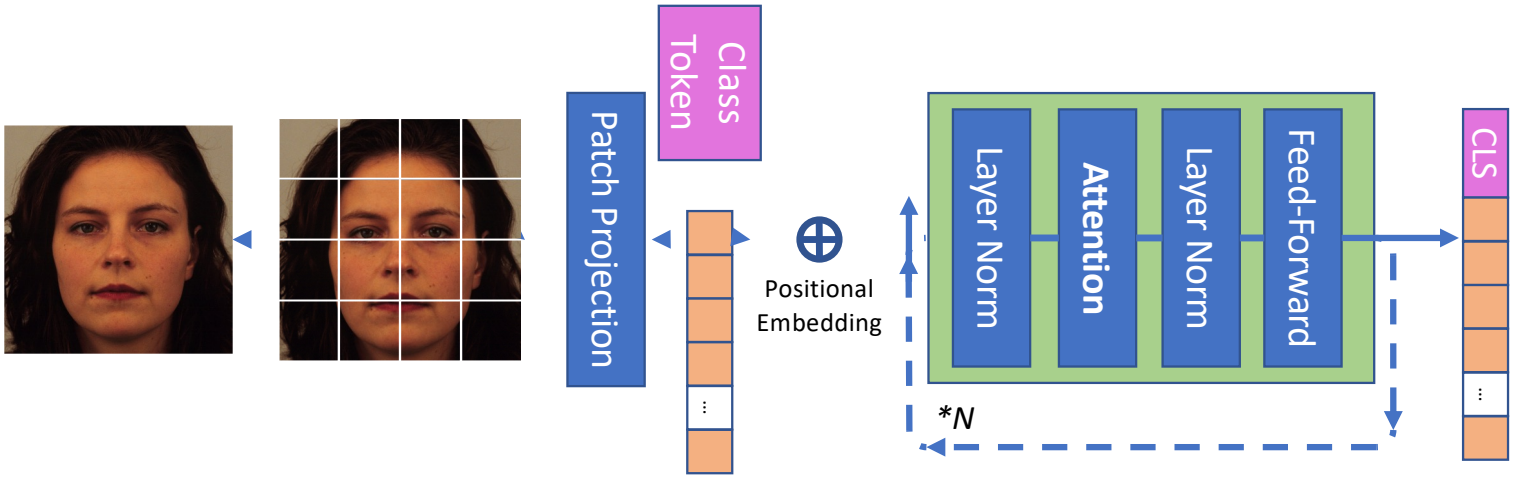
  An input is modelled as the set of pairwise interactions between tokens.

  Tokens in vision can be small image patches.
- **MHSA** can be parallelized, offering faster training… but quadratic operation.

  Methods have been proposed to address this e.g. **SWIN transformer** [5].

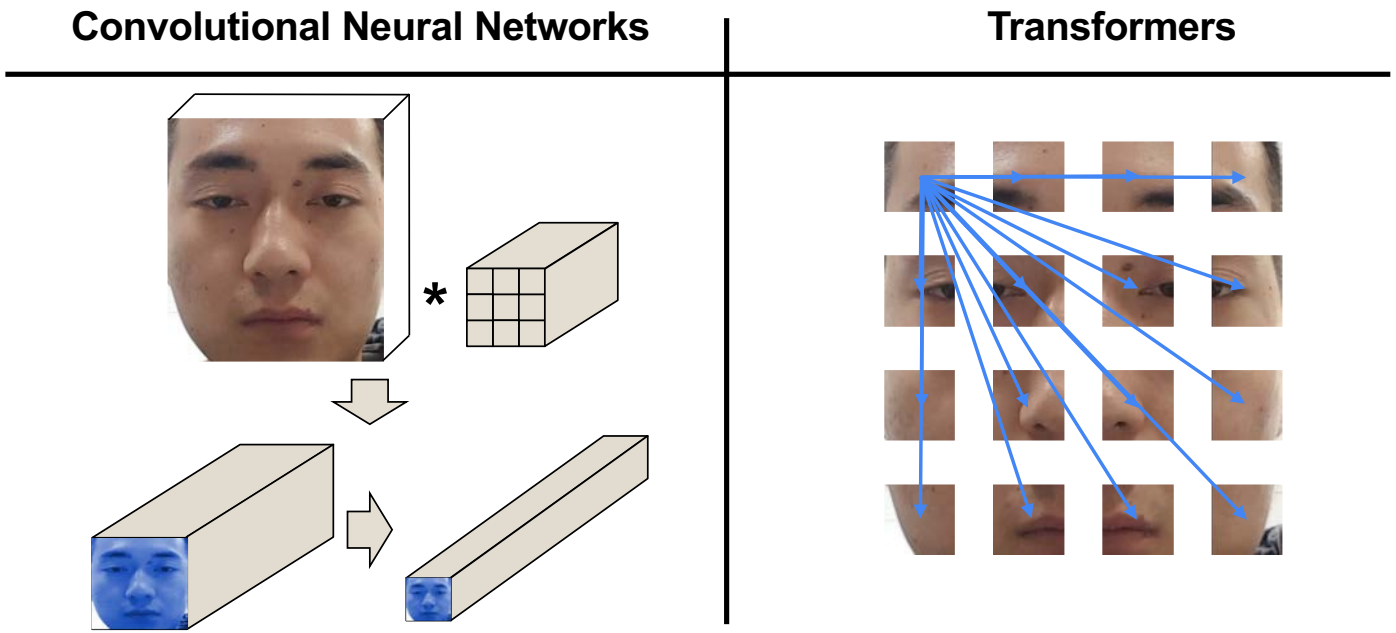
1. Dosovitskiy, Alexey, et al. "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale." International Conference on Learning Representations. 2020.
2. Zhong, Yaoyao, and Deng, Weihong. "Face Transformer for Recognition." arXiv e-prints. 2021.
3. Mao, Jiageng, et al. "Voxel transformer for 3d object detection." Proceedings of the IEEE/CVF International Conference on Computer Vision. 2021.
4. George, Anjith, and Sébastien Marcel. "On the effectiveness of vision transformers for zero-shot face anti-spoofing." 2021 IEEE International Joint Conference on Biometrics (IJCB). IEEE, 2021.
5. Liu, Ze, et al. "Swin transformer: Hierarchical vision transformer using shifted windows." Proceedings of the IEEE/CVF International Conference on Computer Vision. 2021.

Vision Transformers



- Vision Transformers (ViT) use Self-Attention to form a description of an input.
- Huge number of variants. (ViT, DeIT, SWIN, PViT, BEIT, iGPT etc…).
- Have strong general representative capability.
  - Models pre-trained on general data are efficiently transferable to other tasks.
  - Large amount of data required for pre-training.

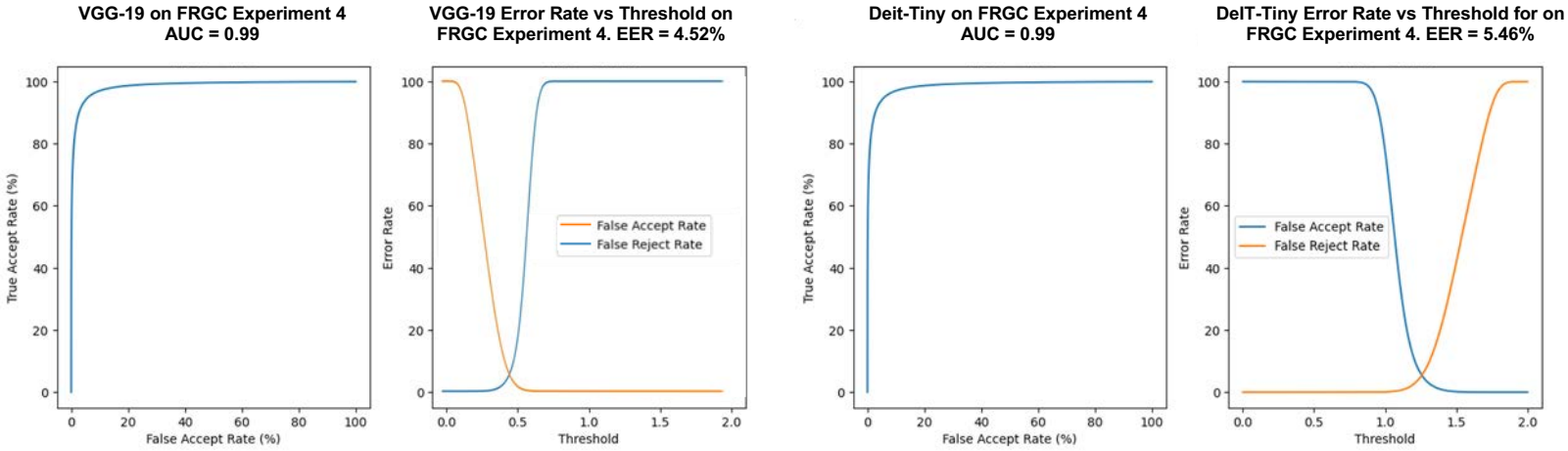
Vision Transformers



- Convolutions in **CNNs** perform feature extraction and aggregation.
  - Local receptive fields (limited by kernel size).
- Attention in **Transformers** performs feature extraction and comparison.
  - Supercharges the features with higher information content.
  - Global receptive fields (every patch communicates with every other patch)

Face Recognition with Transformers

- Data Efficient Image Transformer (DeIT) with orders of magnitude less parameters can perform similarly to a much larger VGG-19 CNN, with only basic transfer learning.
- DeIT-Tiny pre-trained on Imagenet-1k, transfer learned on FRGC experiment 4[1].



143.7M Parameters **5.5M Parameters**

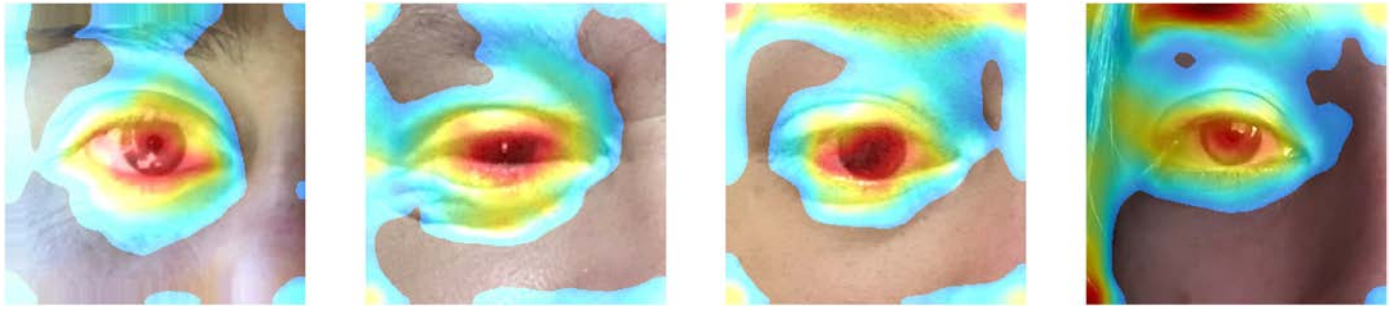
[1] Research conducted for the "Secure Passwordless Authentication for DigitAI identities" project (SPADA)

Transformers Visual Attention

Vision Lab

Face Recognition (FRGC)

Periocular Recognition (UFPR-Periocular)



Humans vs Visual Transformers

- There is a strong correlation between human fixations and transformer attention

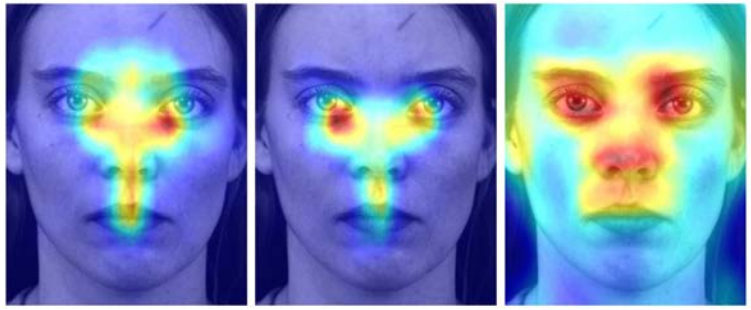


Figure 6. From left to right: fixation densities of male observers, of female observers and ViT attention maps.

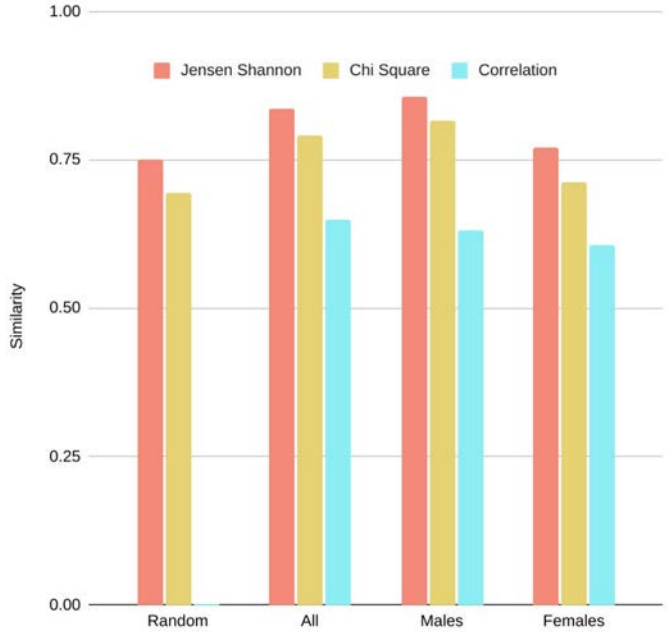


Figure 5. Human vs Transformer attention, split by sex of observer.

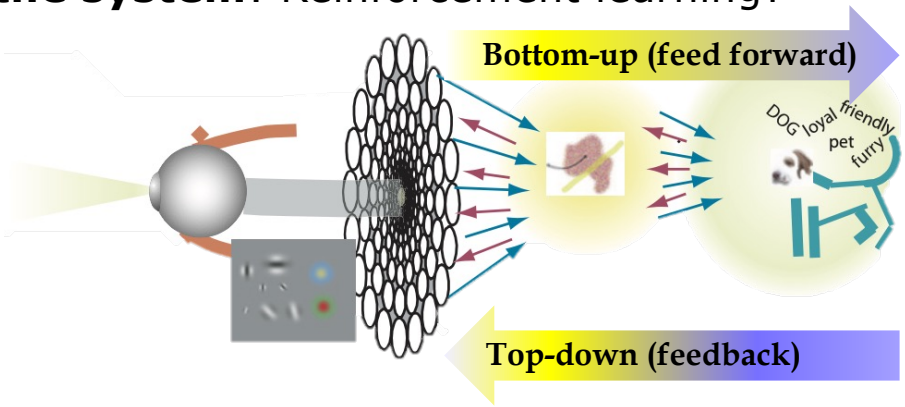
M. Cadoni, A. Lagorio, E. Grosso, T. Jia Huei, C. Chee Seng (2021) "**From early biological models to CNNs: do they look where humans look?**", 25th Int.l Conference on Pattern Recognition ICPR 2020, pp. 6313-6320. doi: 10.1109/ICPR48806.2021.9412717.

Conclusion

- **Deep neural architectures** provide today the current state of the art performance of face recognition *in the wild.*
  - The large number of layers requires a huge amount of data for training to reach a stable configuration of the neural connectivity.
  - They are sensitive to unexpected changes in the spatial frequencies of the input patterns.
- **Simple biologically-inspired networks** may allow to perform very complex visual tasks.
- In biological systems **attention** drives **recognition**.
  - A space-variant **scale-space decomposition** of the input signal allows to select the most informative data.
- The **S1C1** neural architecture, derived from the ***HMAX*** model, with face quality, **outperforms the deep VGG model**.
  - The **peripheral area of the face** (face outline and hair dressing) proved to be very distinctive for recognition.

What about the future?

- **Learn more from biological neural architectures to build network models:** Beyond the retino-cortical topological mapping
- **Learn from human perceptual behaviors:** Improve attention mechanisms; make networks more *curious*
- **Change the learning paradigm:** Exploit interactions; incremental and continuous learning
- **Adversarial attacks and robustness**: Interpolation/ approximation mistakes? How do they compare to optical illusions?
- **Add feedback to the system**: Reinforcement learning?



THANK YOU FOR YOUR