Video Analysis with Deep Learning and without it

Pr. Jenny Benois-Pineau
LABRI UMR 5800/Université Bordeaux

Lecture 2

Visual Attention and its models

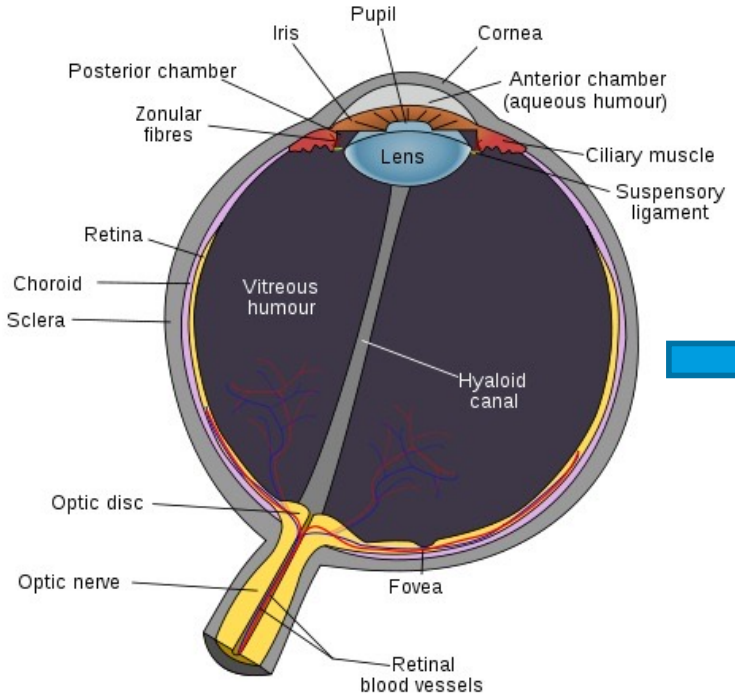
université de BORDEAUX

Visual Attention in Video scene recognition

Summary.

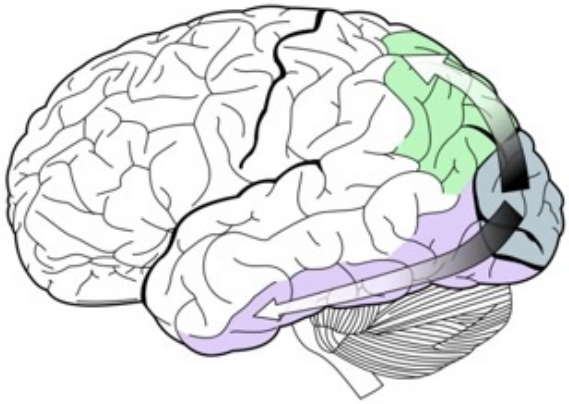
1. Measuring visual attention in images and video
2. Models of visual attention

1. Measuring visual attention in images and videos



Eye

Optic chiasm



Visual cortex

Fovea: central retinal part where vision is the most accurate. Contains cone photoreceptors

[Hubel 95] David H. Hubel. Eye, Brain, and Vision. W. H. Freeman, 2nd edition, may 1995.
[Hérault 01] Jeanny Hérault. De la rétine biologique aux circuits neuromorphiques. Les systèmes de vision. J.M Jolion, hermès edition, 2001.

Bottom-up vs Top Down

➔ Bottom-up

Rapid, unconcious,
attraction by singularities

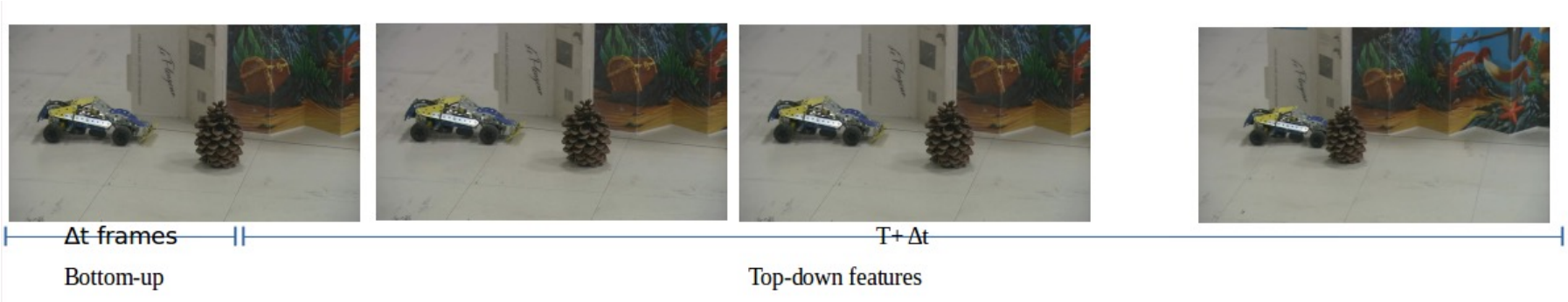
A. M. Treisman and G. Gelade. A feature-integration theory of attention. Cognitive Psychology, 12(1):97-136, 1980

Top-Down

Guided by a visual task

Bottop-up vs Top down

➔ In video : changing in time



In free viewing conditions

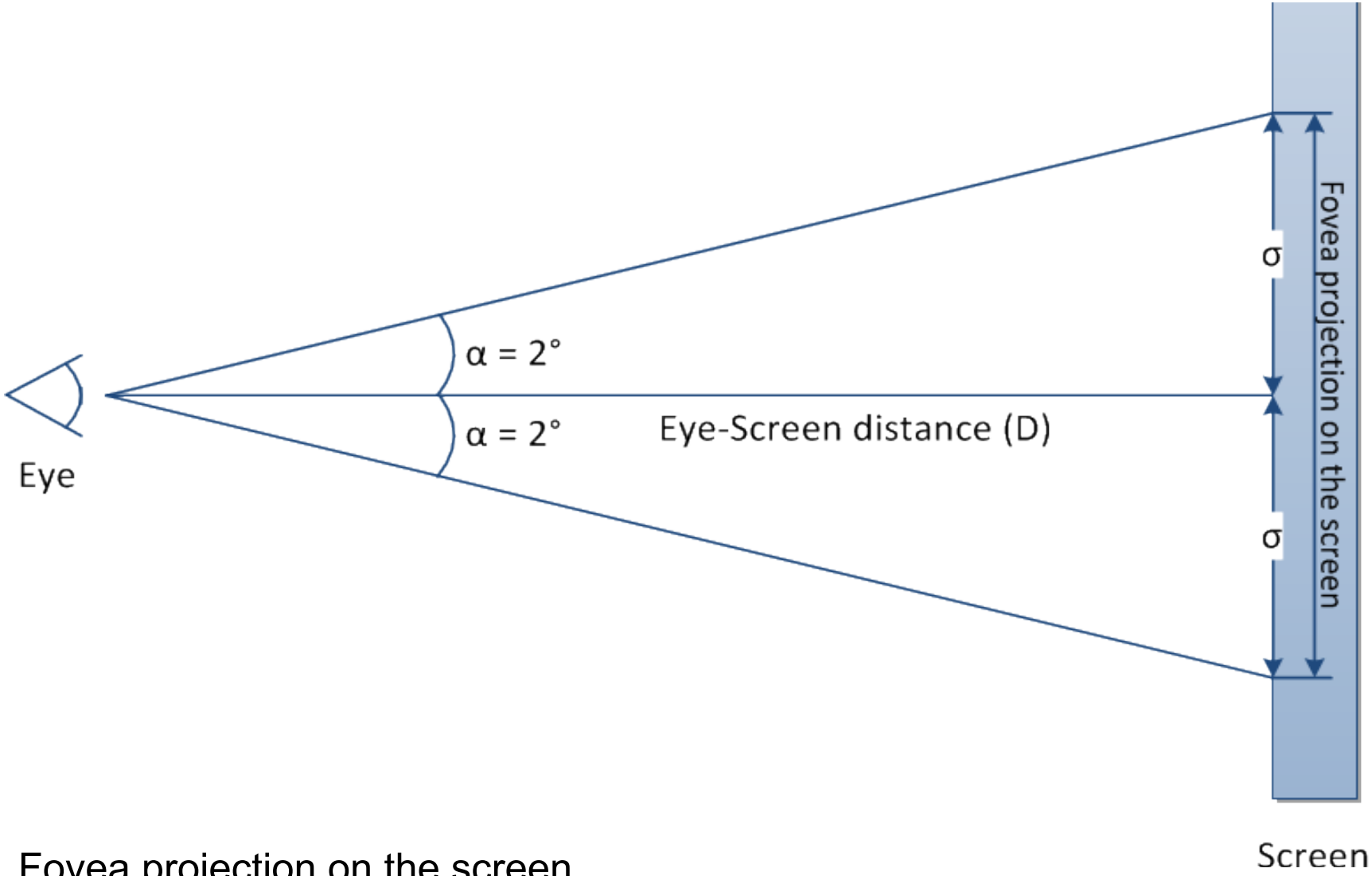
Measuring visual attention

- Gaze Fixation Density Maps (Wooding maps)

$$S_g(X) = [\frac{1}{N_{obs}} \sum_{i=1}^{N_{obs}} (\sum_{m=1}^{M_{fix}} \delta(X - x_{f(m)}))] * G_\sigma(X)$$

- X : spatial coordinates
- $X_{f(m)}$: spatial coordinates of m^{th} visual fixation
- M_{fix} : the number of visual fixations of i^{th} subject
- N_{obs} : number of subjects
- $\delta(.)$: Kronecker symbol

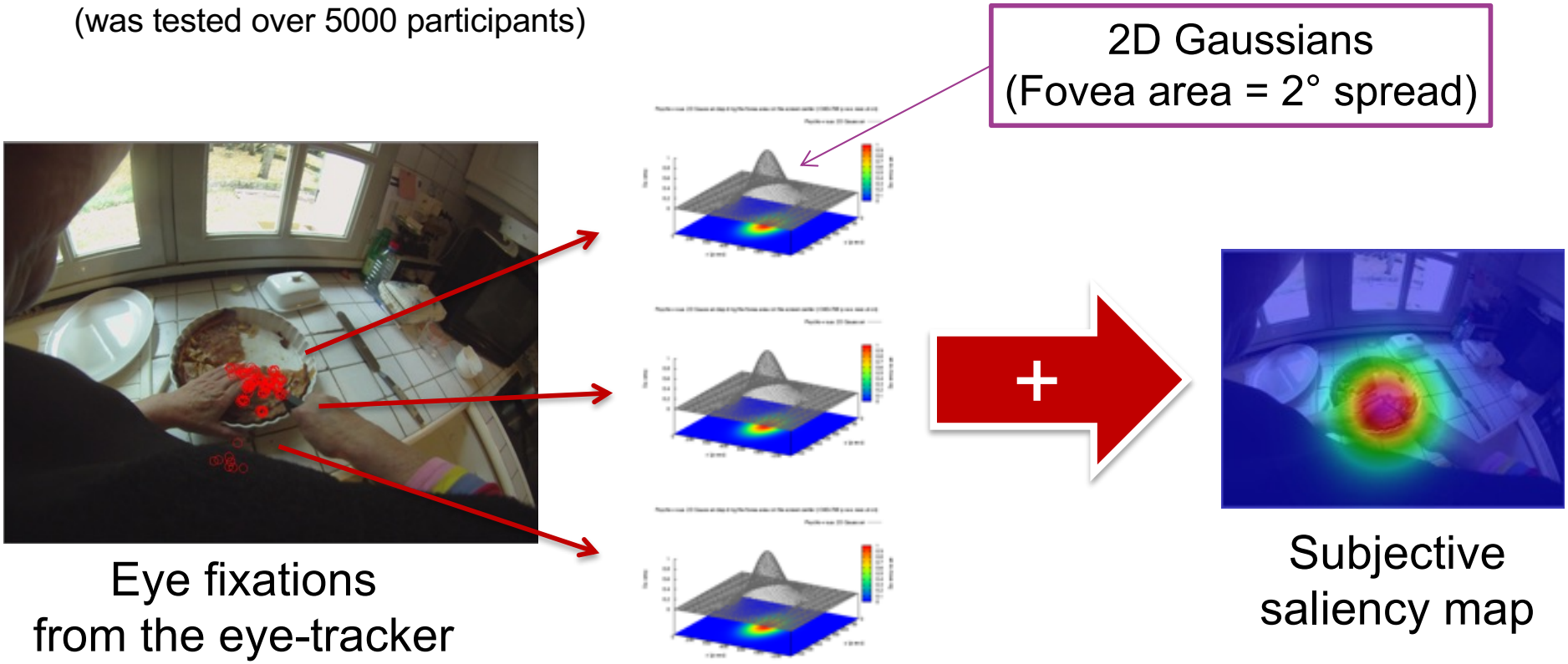
Subjective Saliency – Visual attention map



Fovea projection on the screen

SUBJECTIVE SALIENCY

› D. S. Wooding method, 2002
(was tested over 5000 participants)

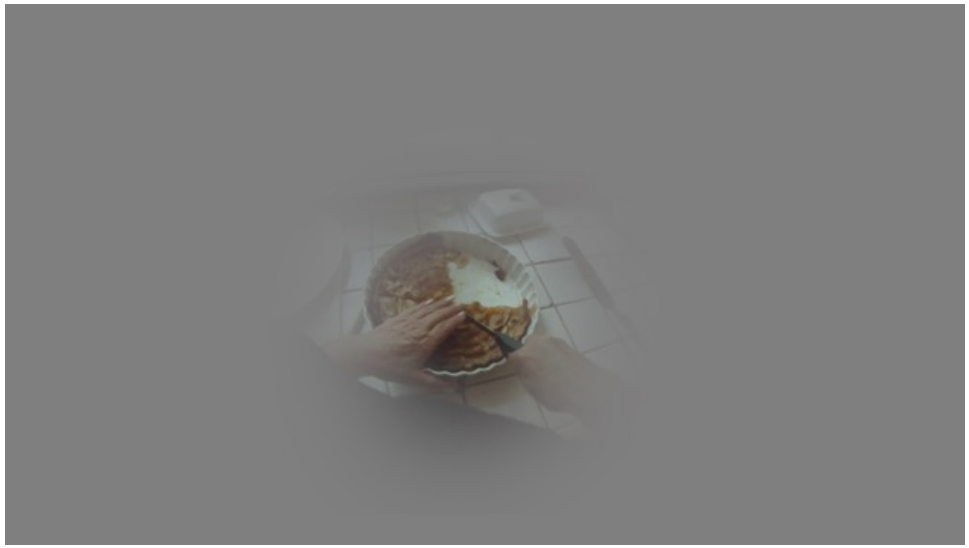


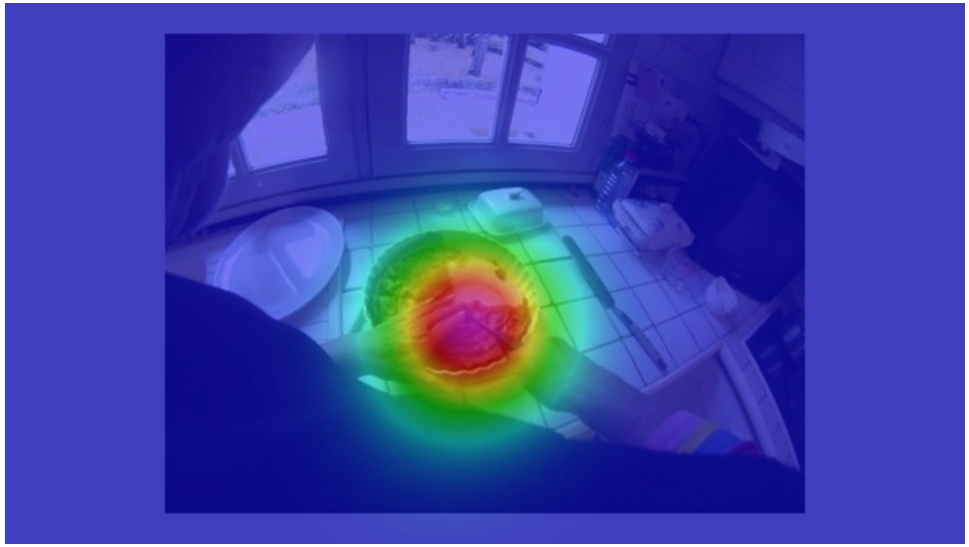
A fixation point indicates the highest resolution region of the image and corresponds to the center of the eye's retina, the fovea. Free viewing conditiond

PSYCHO-VISUAL EXPERIMENT

- Psycho-visual experiment with free viewing conditions
- Gaze measure with an Eye-Tracker (Cambridge Research Systems Ltd. HS VET 250Hz)
- 31 HD video sequences from IMMED database.
- Duration 13’30”
- 25 subjects (5 discarded)
- 6 562 500 gaze positions recorded

SUBJECTIVE SALIENCY

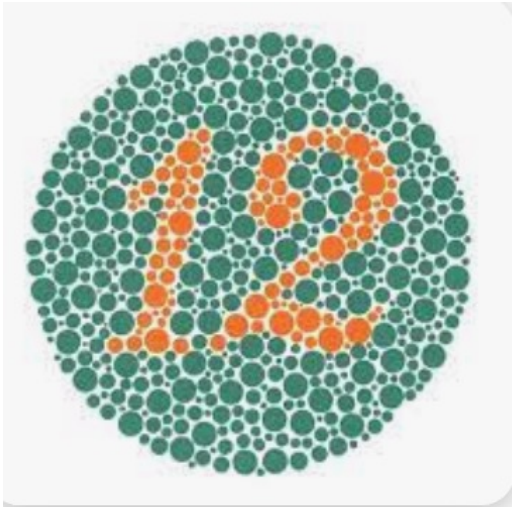




Task-Driven Psycho-visual experiment

- **Problem : recognition of architectural styles of Mexican Buildings**

- **Protocol :**
- written instructions to participants;
- Ishihara test (detection of color visión anomalies)



- **Total images:** **284**
-
- Time for image displaying: 3 seconds
- Time for gray frame displaying: 1 second
- Time for calibration: 60 seconds
- Time to read instructions: 180 seconds

Abraham Montoya Obeso, Jenny Benois-Pineau, Mireya Saraí García-Vázquez, Alejandro Alvaro Ramírez-Acosta: Visual vs internal attention mechanisms in deep neural networks for image classification and object detection. Pattern Recognit. 123: 108411 (2022)

Example of Saliency maps from Task-driven visual experiment

Mexculture284 dataset, 142 buildings, Fixations from 23 participants, Visual task : recognition of architectural styles. Available at https://api.nakala.fr/data/11280%2F5712e468/1e412e0a43b5635365293b249feb9d53d74b5dc8, https://www.labri.fr/projet/AIV/MexCulture142.php

Some hints to explanation

- Guy Buswell : *How people look at pictures* (1935), Univ. Chicago Press
- A. L. Yarbus, *Eye Movements and Vision (*1967). New York: Plenum Press,
- Important contribution (1): cognitive factors such as viewer's task can have a strong effect upon how a picture is inspected.

- Important contribution (2): central bias hypothesis
- - at the beginning of the observation subject look in the center of the picture.

Modeling Visual Attention

- Several approaches
  - Bottom-up or top-down
  - Overt or covert attention
  - Spatial or spatio-temporal
  - Scanpath or pixel-based saliency

- Features
  - Intensity, color, and orientation (Feature Integration Theory [1]), HSI or L*a*b* color space
  - In video : Relative motion [2]

- Plenty of models in the literature
  - Classical (Feature integration theory)
  - Deep

[1] Anne M. Treisman & Garry Gelade. A feature-integration theory of attention. Cognitive Psychology, vol. 12, no. 1, pages 97–136, January 1980.
[2] Scott J. Daly. Engineering Observations from Spatiovelocity and Spatiotemporal Visual Models. In IS&T/SPIE Conference on Human Vision and Electronic Imaging III, volume 3299, pages 180–191, 1 1998.

Comaparison metrics for visual attention maps

- NSS

$$NSS = \frac{\overline{S_{subj} \times S_{obj}^{N}} - \overline{S_{obj}}}{\sigma\left(S_{obj}\right)}$$

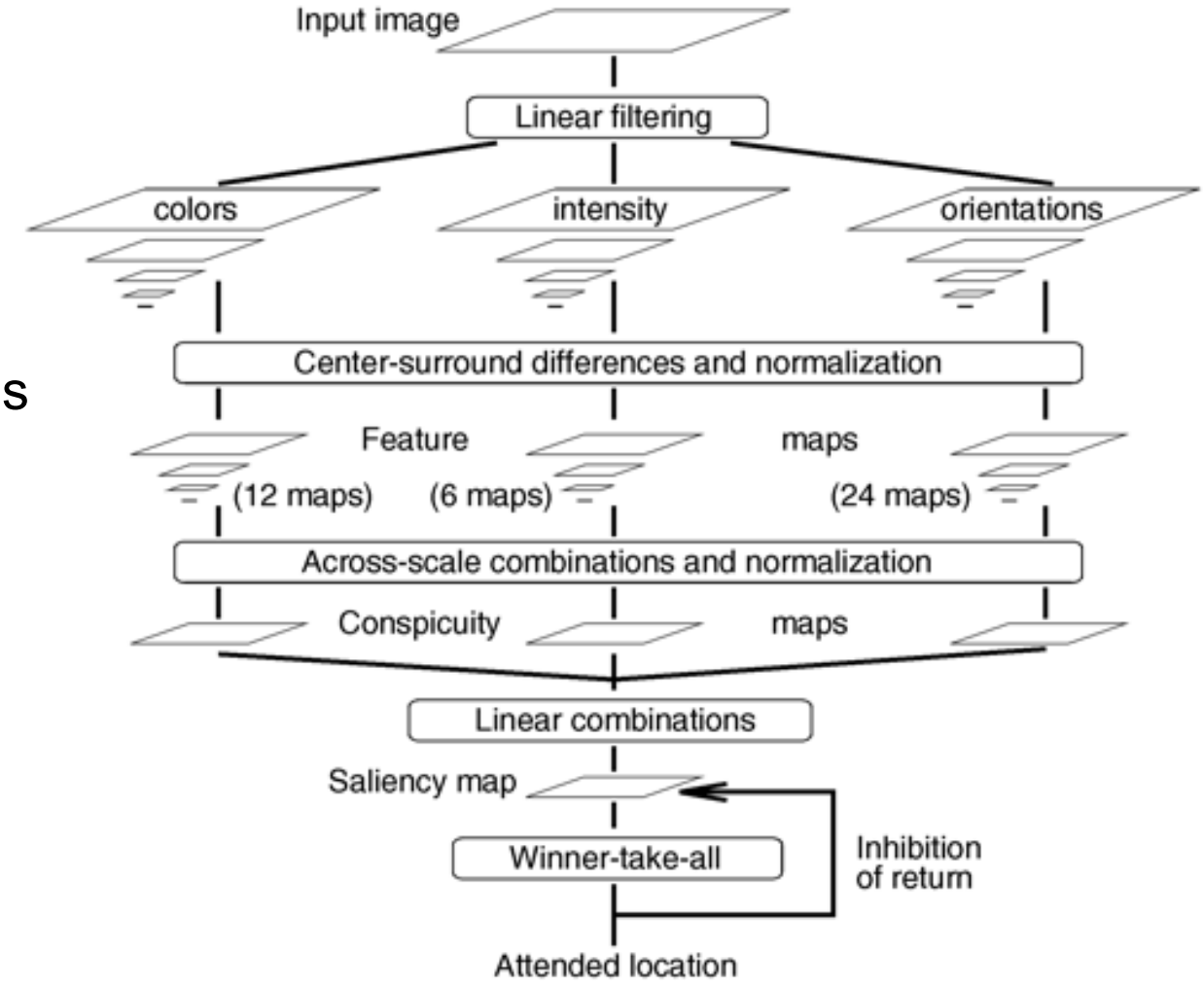
- PCC

$$r\left(S_1, S_2\right) = \frac{\mathrm{cov}\left(S_1, S_2\right)}{\sigma\left(S_1\right) \cdot \sigma\left(S_2\right)}$$

- AUC : thresholding of S_{subj} and S_{obj} Then plotting teh ROC curve :
- TPR= TP/(TP+FN) against FPR = FP/(TP+FN)

ITTI'S MODEL

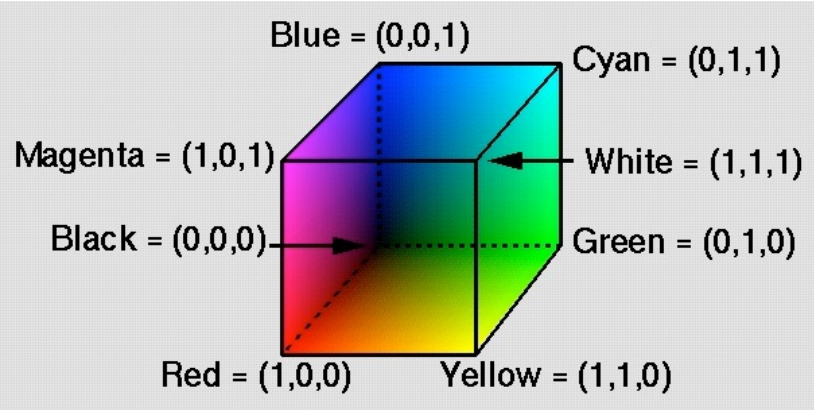
- The First most popular model
- Designed for still images
- Takes into account sensitivity to contrats, colours, orientations
- Does not consider the temporal dimension of videos



[1] Itti, L.; Koch, C.; Niebur, E.; , "A model of saliency-based visual attention for rapid scene analysis , » *Pattern Analysis and Machine Intelligence, IEEE Transactions on* , vol.20, no.11, pp.1254-1259, Nov 1998

Itti's model. Early Visual features (1)

(1) Image transformation/linear filtering : goal – to transform into the colour system more adapted to human perception. Initial image is in r,g,b system



« uniform » system

$$I = (r+g+b)/3$$

$$R = r-(g+b)/2$$

$$G = g-(r+b)/2$$

$$B = b-(r+g)/2$$

$$Y = (r+g)/2-|r-g|/2$$

Itti's model. Early visual features(2)

- (2) Gaussian Pyramids :

$$I * G \downarrow, \qquad G(x,y,\sigma) = \frac{1}{2\pi\sigma^2} e^{-\left(\frac{x^2+y^2}{2\sigma^2}\right)}$$

- scales : 0,…,8 on I, R, G, B, Y

- (3) « Center-surround difference » :
- base scale:
-
- neighbouring scale :
- CSD's are computed accros scales

$$c \in \{2,3,4\} \qquad s = c + \delta, \qquad \delta \in \{3,4\}$$

Itti's model. Early visual features (3)

- **Intensity contrast** : $I(c,s) = \left| I(c) \div I(s) \right|$
- 6 maps
- **Colour contrast** :
- 12 maps

$$RG(c,s) = \left| (R(c) - G(c)) \div (G(s) - R(s)) \right|$$

$$BY(c,s) = \left| (B(c) - Y(c)) \div (B(s) - Y(s)) \right|$$

- - operations on interpolated images

$\div, \oplus$

Itti's model. Early visual features (4)

➔ **Orientation :**

➔ Gabor pyramids:

$$h(x,y) = \exp\left(-\frac{x_0^2 + \gamma y_0^2}{2\sigma^2}\right)\cos\left(\frac{2\pi}{\lambda}x_0\right)$$

$$x_0 = x\cos(\theta) + y\sin(\theta),$$

- impulse response of the Gabor Filter. $O(\theta,\sigma)$

$$y_0 = -x\sin(\theta) + y\cos(\theta)$$

- Center – surround :

➔24 maps.

➔

$$\theta \in \{0°, 45°, 90°, 135°\} \quad \sigma \in [0,...,8]$$

$$O(c,s,\theta) = \left|O(c,\theta) \div O(s,\theta)\right|$$

Itti's model. Normalisation of maps(5)

➔ Feature maps represent non comparable modalities.

- Normalizing in a range [0...1] ;
- Finding a location of a map's global maximum
- Computing the average of local maxima $[0,...,M]$
- Multiplying the map by $(M-\overline{m})^2$

➔(to stress the most active location, if the difference is small, the map is suppressed) M $\overline{m}$

➔This coarsly replicates cortical lateral inhibition mechanisms: neighbouring similar features inhibit each other.

Itti's model. Conspicuity maps

- All maps are combined at the scale 4 via –
  - interpolation of each map at scale 4
  - pixel-by pixel adition
- Intensity map:

$$\bar{I} = \oplus_{c=2}^{c=4} \oplus_{s=c+3}^{s=c+4} N(I(c,s))$$

- Colour map:

$$\bar{C} = \oplus_{c=2}^{c=4} \oplus_{s=c+3}^{s=c+4} (N(RG(c,s) + BY(c,s))$$

- Orientation map:

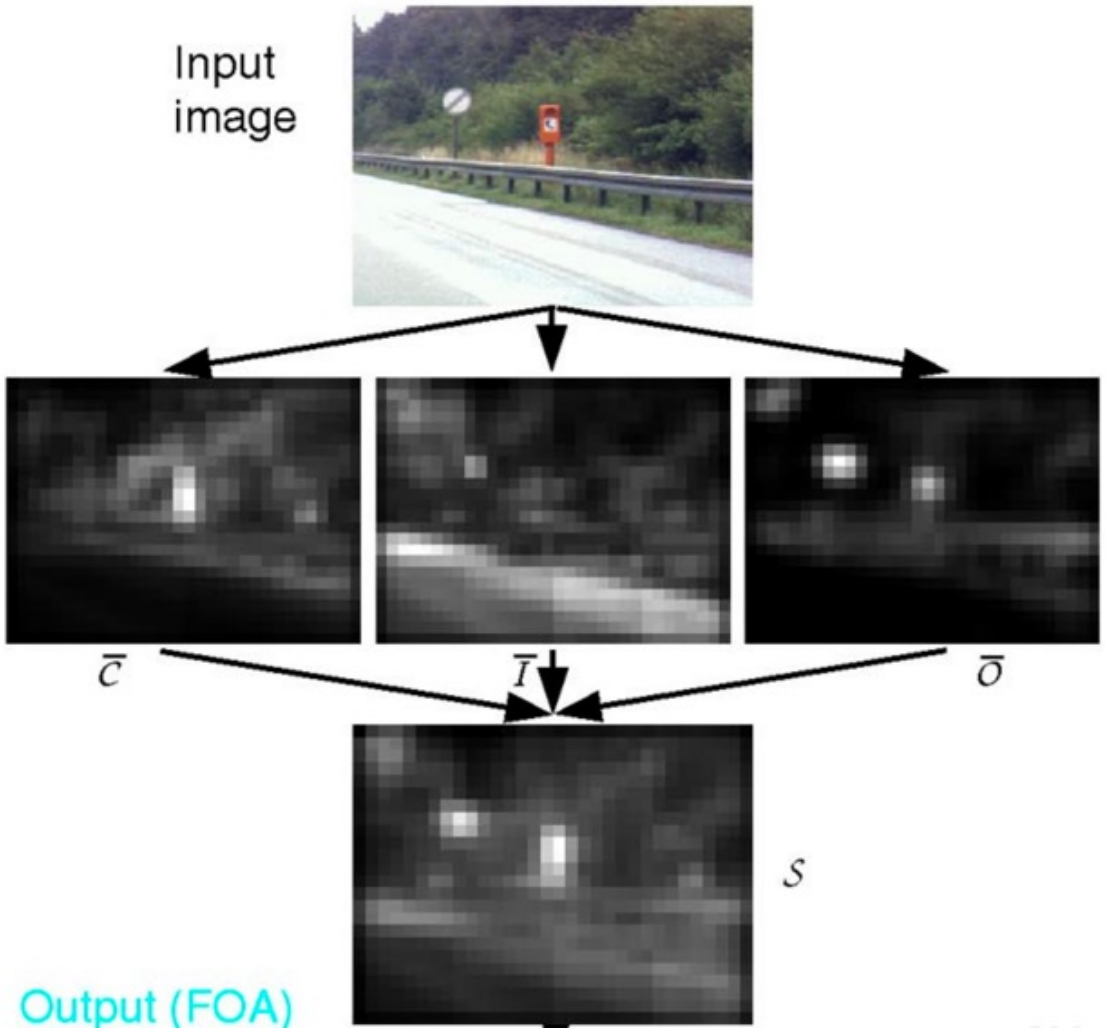
$$\bar{O} = \sum_{\theta=\{0°,45°,90°,135°\}} N(\oplus_{c=2}^{c=4} \oplus_{s=c+3}^{s=c+4} (N(O(c,s,\theta)))$$

Itti's model. Saliency map

$$S = \frac{1}{3}(N(\bar{I}) + N(\bar{C}) + N(\bar{O}))$$

- At any given time the maximum of Saliency map defines the most salient image location to which the focus of attention is directed.
- Modeling by a neuronal network « winner takes all ».

Combination of Intensity, Color and Orientation cues



From Laurent Itti, Christof Koch & Ernst Niebur. A Model of Saliency-Based Visual Attention for Rapid Scene Analysis. IEEE Trans. Pattern Anal. Mach. Intell., vol. 20, no. 11, pages 1254–1259, November 1998.

Example. Itti's model on a video frame

Saliency in Video

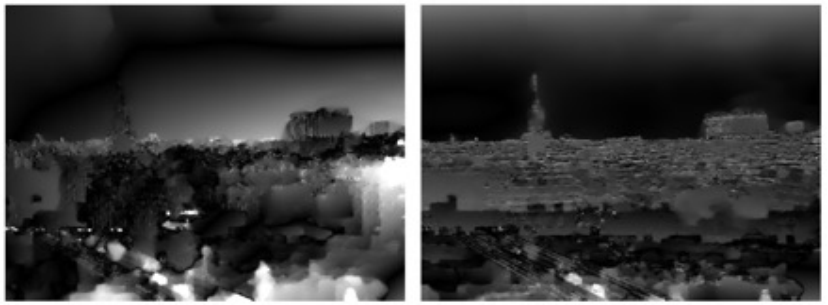
- ➔ Sensitivity to residual motion
- ➔ Example: free viewing conditions

Free viewing conditions : cognitive influence : anticipation, semantic elements, recognition (EU Dem@care demo).

Saliency in Video; Prediction in the era of Deep Learning

➔ Résidual motion

$$\overrightarrow{M_r}(x,y) = \overrightarrow{M_\theta}(x,y) - \overrightarrow{M_c}(x,y)$$



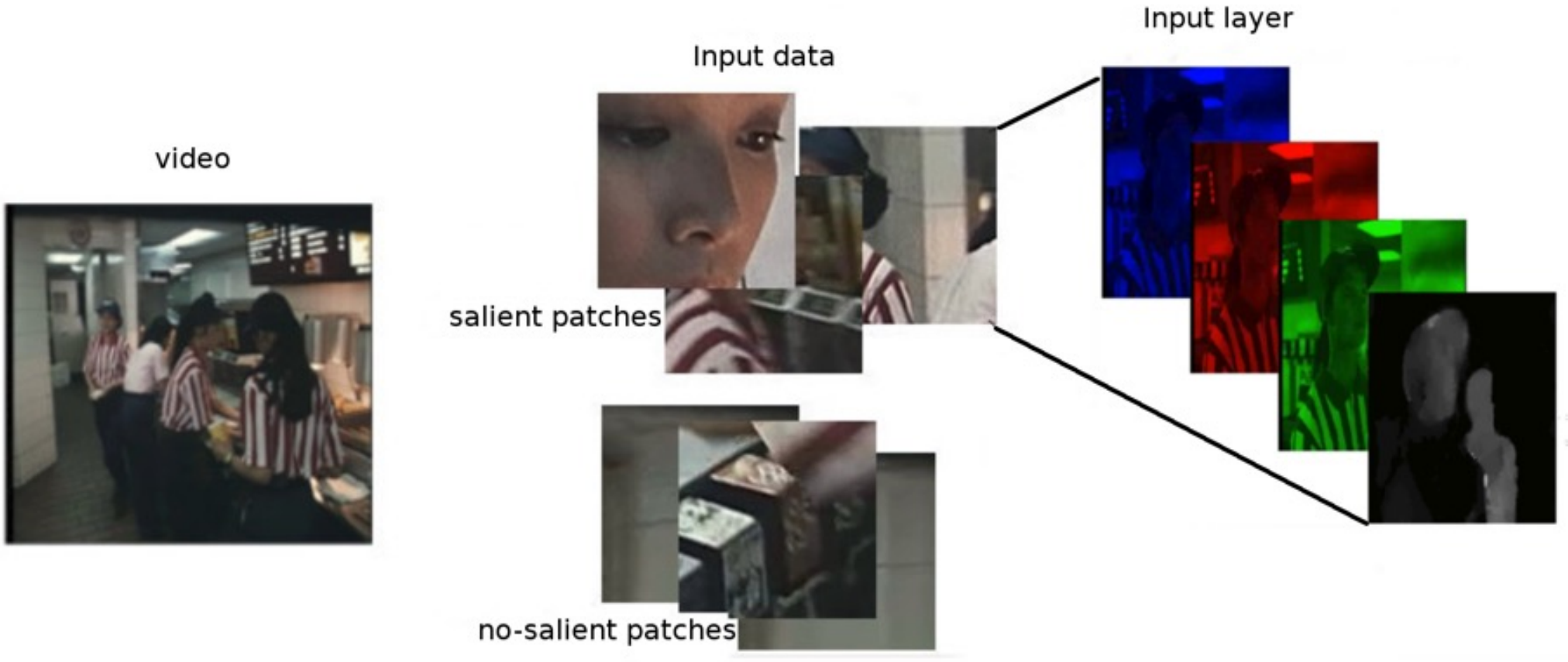
➔ Normalization

$$f^{mot}(I,(x,y)) = \frac{\| \overrightarrow{M_r}(x,y) \|_2^2}{\max_{(x,y)in\Omega} \| \overrightarrow{M_r}(x,y) \|_2^2}$$

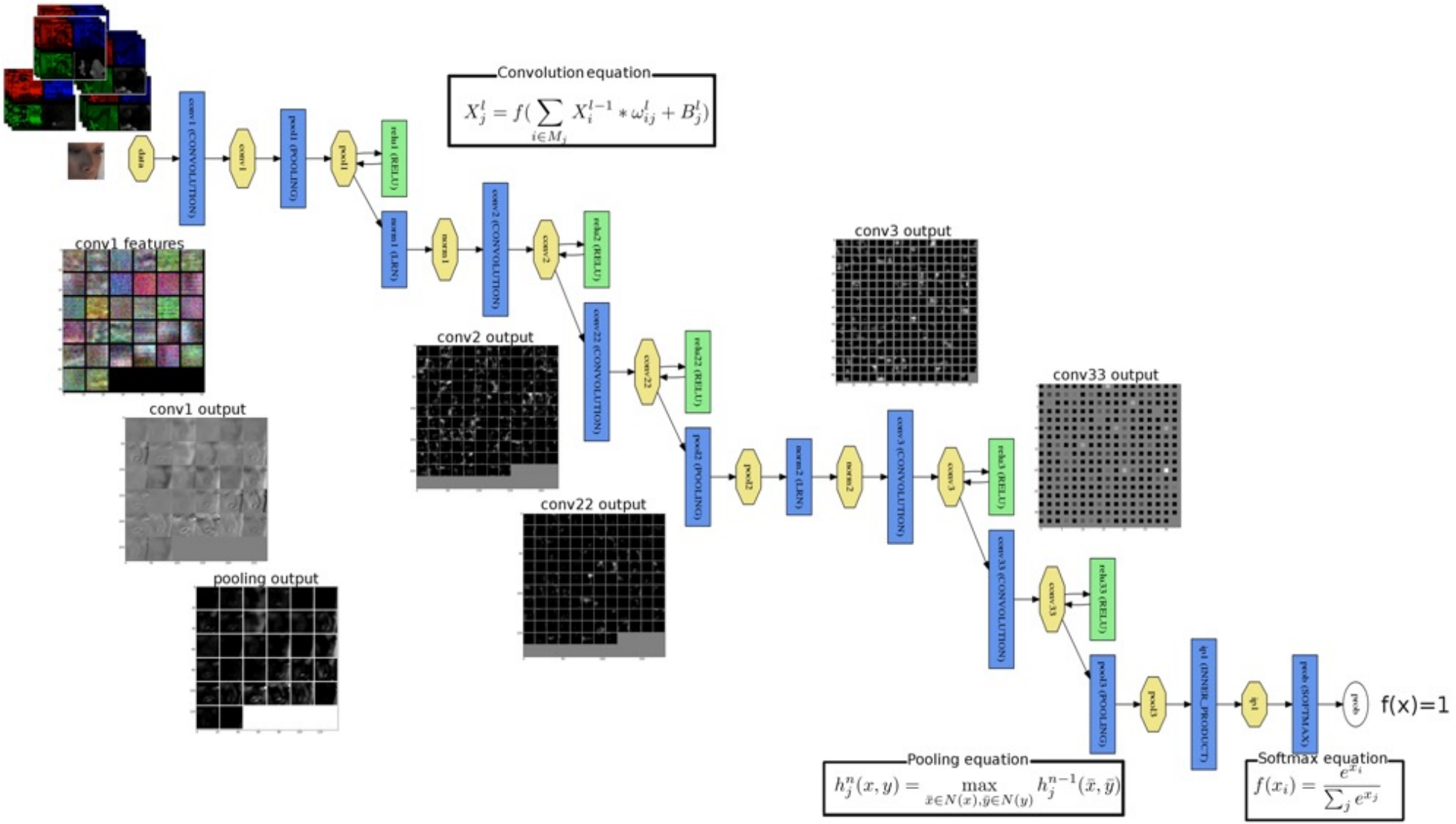
dx dy

Fusion of RGB and Motion Infromation

- ➔ Early fusion is performed in a data space.
- ➔ Example : saliency prediction : RGB + Residual Motion



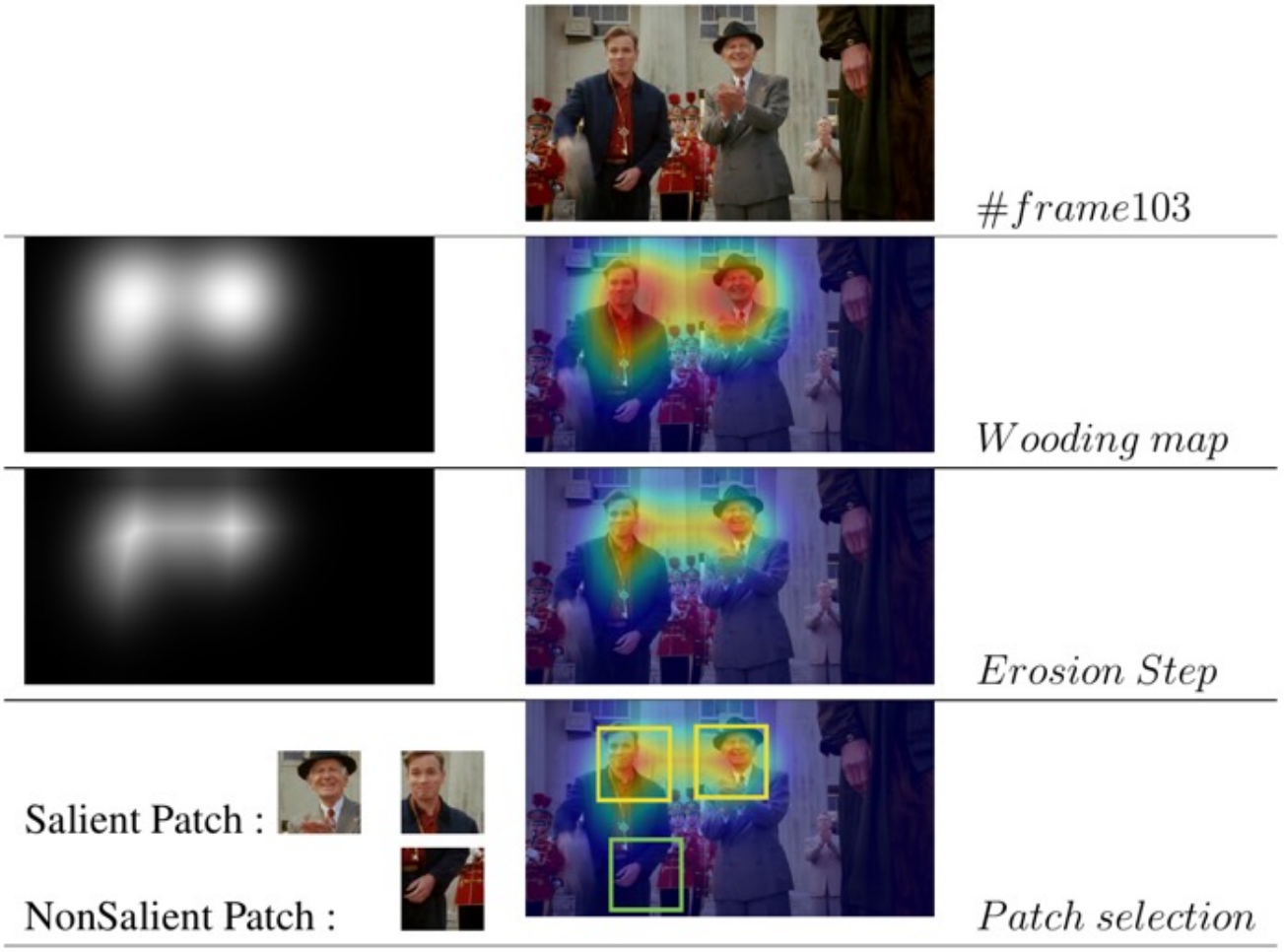
Architecture: AlexNet-like deep network "Chabonet"



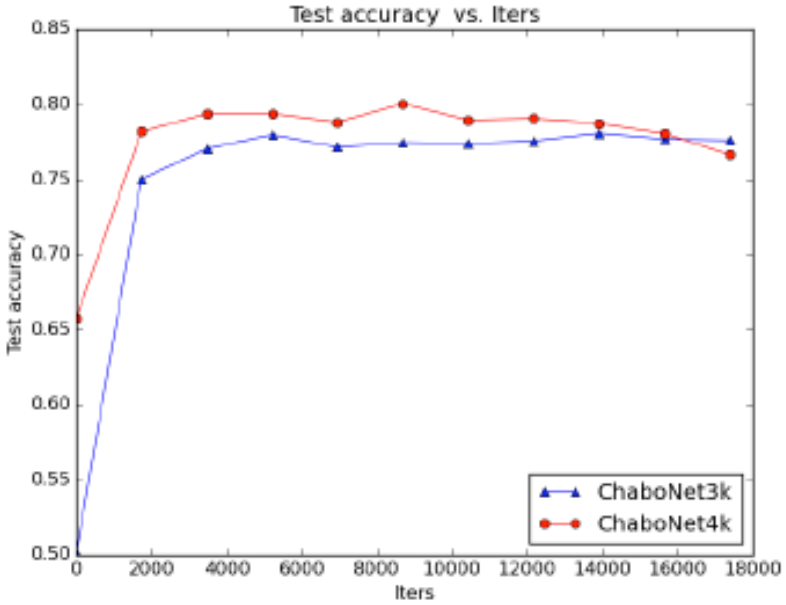
Prediction of patches' class and interpolation

Ground Truth for training

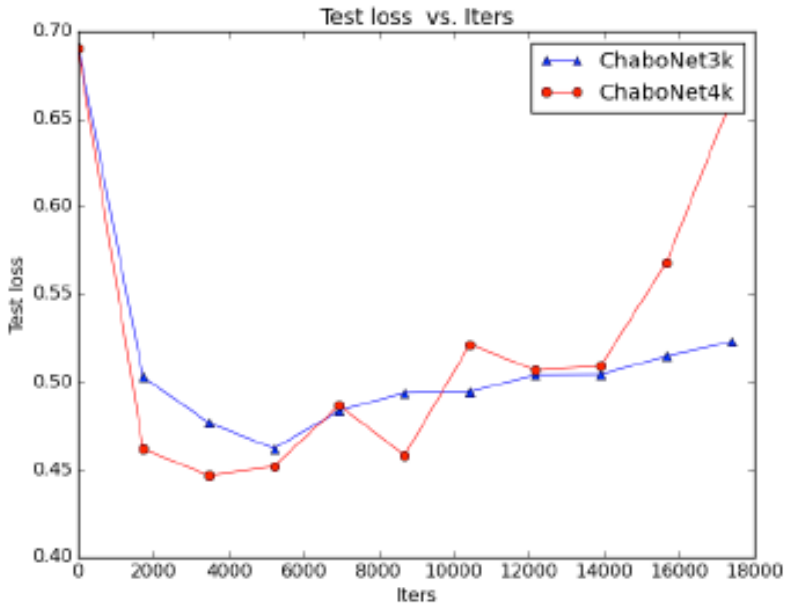
Principe: A salient patch P_i is selected on the basis of the GFDM around local maximums



Results of early fusion aproach



(a) Accuracy vs iterations

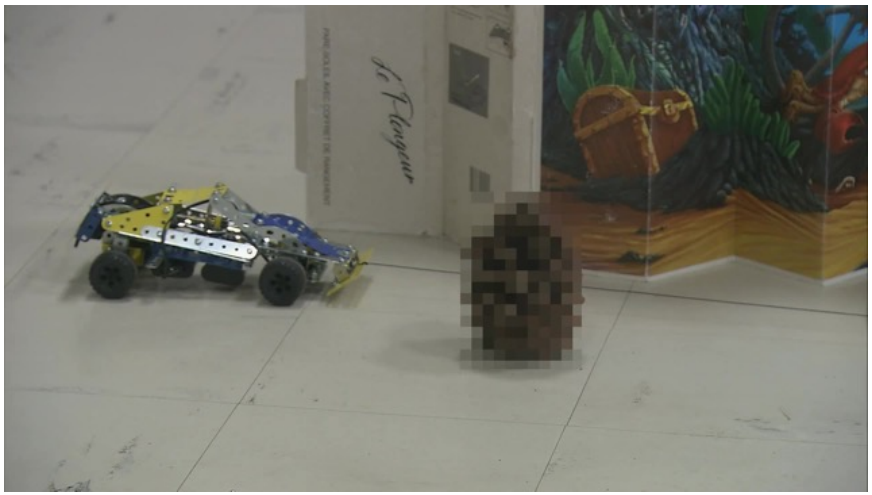


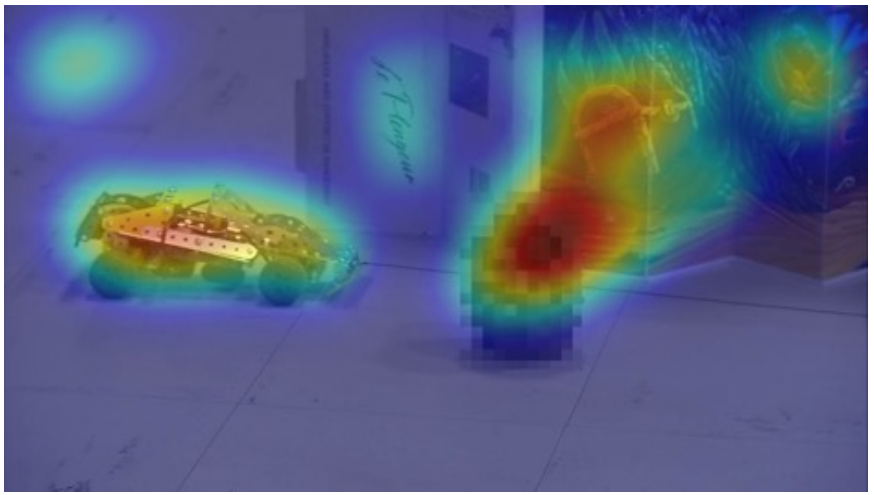
(b) Loss vs iterations

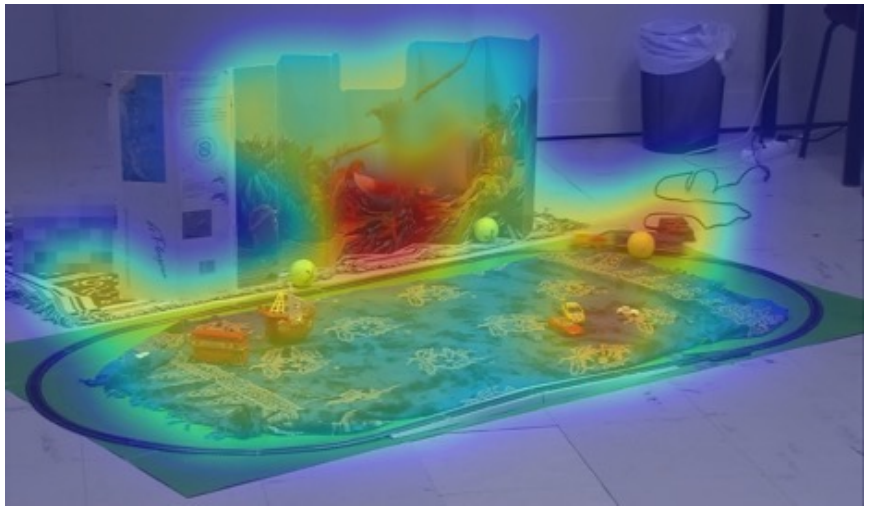
	$ChaboNet3k$	$ChaboNet4k$
$min_{(\#iter)}$	$50.11\%_{(\#0)}$	$65.73\%_{(\#0)}$
$max_{(\#iter)}$	$77.98\%_{(\#5214)}$	$80.05\%_{(\#8690)}$
$avg \pm std$	$77.30\% \pm 0.864$	$78.73\% \pm 0.930$

(c)The accuracy results on HOLLYWOOD dataset

Examples of results on specifically degraded videos







What will we do in the project(1)?

➔ Working on Egocentric video with, ego eye-tracker

Grasping-in-the-Wild (LABRI)
Dataset Available at CNRS Nakala
https://www.labri.fr/projet/AIV/graspinginthewild.php

I. González-Díaz, J. Benois-Pineau, J.-Ph. Domenger, D. Cattaert, A. de Rugy: Perceptually-guided deep neural networks for ego-action prediction: Object grasping. Pattern Recognit. 88: 223-235 (2019)

GTEA Dataset, Georgia Tech

A. Fathi, X. Ren and J. M. Rehg, "Learning to recognize objects in egocentric activities," *CVPR 2011*, 2011, pp. 3281-3288, doi: 10.1109/CVPR.2011.5995444.