Etude et prédiction d'attention visuelle avec les outils d'apprentissage profond en vue d'évaluation des patients atteints des maladies neuro-dégénératives

Souad Chaabouni

THÈSE EN COTUTELLE PRÉSENTÉE
POUR OBTENIR LE GRADE DE

DOCTEUR DE

L'UNIVERSITÉ DE BORDEAUX

ET DE L'UNIVERSITÉ DE SFAX

.....

ÉCOLE DOCTORALE DE MATHÉMATIQUES ET D'INFORMATIQUE
SPÉCIALITÉ Informatique
Par **Souad, CHAABOUNI**

Étude et prédiction d'attention visuelle avec les outils d'apprentissage profond en vue d'évaluation des patients atteints des maladies neuro-dégénératives

.....

Sous la direction de Pr. Jenny BENOIS-PINEAU et de Pr. Chokri BEN AMAR.
Soutenue le . . . 08/12/2017

Membres du jury :

Pr. Le Callet, Patrick	Professeur, Polytech Nantes-Nantes University	Président
Pr. Precioso, Frederic	Professeur, University Nice Sophia Antipolis	Rapporteur
Pr. Essoukri-Ben Amara, Najoua	Professeur, University of Sousse	Rapporteur
Pr. Chtourou, Mohamed	Professeur, University of Sfax	Examinateur
Pr. Tison, François	Professeur, Bordeaux University	Invité
Pr. Desbarats, Pascal	Professeur, Bordeaux University	Invité

Titre : Étude et prédiction d'attention visuelle avec les outils d'apprentissage profond en vue d'évaluation des patients atteints des maladies neuro-dégénératives

Résumé : Cette thèse est motivée par le diagnostic et l'évaluation des maladies neurodégénératives et dans le but de diagnostique sur la base de l'attention visuelle. Néanmoins, le dépistage à grande échelle de la population n'est possible que si des modèles de prédiction automatique suffisamment robustes peuvent être construits. Dans ce contexte nous nous intéressons à la conception et le développement des modèles de prédiction automatique pour un contenu visuel spécifique à utiliser dans l'expérience psycho-visuelle impliquant des patients atteints des maladies neurodégénératives. La difficulté d'une telle prédiction réside dans une très faible quantité de données d'entraînement.

Les modèles de saillance visuelle ne peuvent pas être fondés sur les caractérisitiques "bottom-up" uniquement, comme le suggère la théorie de l'intégration des caractéristiques. La composante "top-down" de l'attention visuelle humaine devient prépondérante au fur et à mesure d'observation de la scène visuelle. L'attention visuelle peut être prédite en se basant sur les scènes déjà observées. Les réseaux de convolution profonds (CNN) se sont révélés être un outil puissant pour prédire les zones saillantes dans les images statiques. Dans le but de construire un modèle de prédiction automatique pour les zones saillantes dans les vidéos naturels et intentionnellement dégradées, nous avons conçu une architecture spécifique de CNN profond. Pour surmonter le manque de données d'apprentissage, nous avons conçu un système d'apprentissage par transfert dérivé de la méthode de Bengio. Nous mesurons ses performances lors de la prédiction de régions saillantes. Les résultats obtenus sont intéressants concernant la réaction des sujets témoins normaux contre les zones dégradées dans les vidéos. La comparaison de la carte de saillance prédite des vidéos intentionnellement dégradées avec des cartes de densité de fixation du regard et d'autres modèles de référence montre l'intérêt du modèle développé.

Mots clés : Réseaux de convolution profond, apprentissage par transfert, vision par ordinateur, modèle de saillance, attention visuelle, maladies neuro-dégénératives, mouvement résiduel, vidéos naturels

INRIA Bordeaux Sud-Ouest-200818243Z Institut de Mathématiques de Bordeaux (IMB)-UMR 5251 Laboratoire Bordelais de Recherche en Informatique (LABRI)-UMR 5800

Title : Study and prediction of visual attention with deep learning networks in view of assessment of patients with neurodegenerative diseases

Abstract :

This thesis is motivated by the diagnosis and the evaluation of the dementia diseases and with the aim of predicting if a new recorded gaze presents a complaint of these diseases. Nevertheless, large-scale population screening is only possible if robust prediction models can be constructed. In this context, we are interested in the design and the development of automatic prediction models for specific visual content to be used in the psycho-visual experience involving patients with dementia (PwD). The difficulty of such a prediction lies in a very small amount of training data.

Visual saliency models cannot be founded only on bottom-up features, as suggested by feature integration theory. The top-down component of human visual attention becomes prevalent as human observers explore the visual scene. Visual saliency can be predicted on the basis of seen data. Deep Convolutional Neural Networks (CNN) have proven to be a powerful tool for prediction of salient areas in static images. In order to construct an automatic prediction model for the salient areas in natural and intentionally degraded videos, we have designed a specific CNN architecture. To overcome the lack of learning data we designed a transfer learning scheme derived from bengio's method. We measure its performances when predicting salient regions. The obtained results are interesting regarding the reaction of normal control subjects against degraded areas in videos. The predicted saliency map of intentionally degraded videos gives an interesting results compared to gaze fixation density maps and other reference models.

Keywords : Deep convolutional networks, transfer learning, computer vision, saliency models, visual attention, neuro-degenerative diseases, residual motion, natural videos

INRIA Bordeaux Sud-Ouest-200818243Z Institut de Mathématiques de Bordeaux (IMB)-UMR 5251 Laboratoire Bordelais de Recherche en Informatique (LABRI)-UMR 5800

Acknowledgements

First I would like to thank my supervisor Prof. Jenny Benois-Pineau and Prof. Chokri Ben Amar for guiding me throughout these three years with their scientific rigor and numerous advices.

I am truly thankful that Prof. Frederic Precioso and Prof. Najoua Essoukri-Ben Amara have accepted to review this manuscript. I also would like to thank the members of the jury Prof. Patrick Le Callet, Prof. Mohamed Chtourou, Prof. François Tison and Prof. Pascal Desbarats to be part of my thesis defense committee.

Regarding my own laboratory, the LaBRI, I would like to thank in particular Dr. Boris Mansencal for his technical advice especially when teaching master level. I would like to thank all the people from the administration and the system team who helped me so many times and definitely contribute to the good mood at the LaBRI, thank you all.

I would like to thank the Linnaeus-university and especially Prof. Andrei Khrennikov for his two-month invitation to the mathematical department, which allowed us to understand the mathematical aspect of deep network.

I would like to thank all phd-students and research engineers even I met in LaBRI or in the conferences. Lamis, Rahma, Manel, Samiha, Pierre-Marie, Mariem, my cousin Nadia and her brother Amine, Miguel, Soufiane, Kilian, Rémi, and Christelle,..., I am very happy to know you and you have a special place in my heart.

I would like to thank all the team of Erasmus Mendus (UnetBa). I met a special friends with more than 25 nationality, Asma, Abir, Nina, Natsuki and Sami, ... I would like to thank you for your support and love.

And finally I cannot finish without expressing all the gratitude I have for my family for their support during these three years but really, simply for all these years of my life, period.

for my parents ***Moufida*** *and* ***Fayçal****,*
for my brother ***Wael****,*
for my sister ***Wiem****,*
for the princesses of the family ***Yassmine*** *and* ***Farah****.*

Contents

List of Figures

List of Tables

Glossary

AUC Area Under the Curve. 42

CAM Class activation map. 67

CNN Convolutional neural network. 25, 41, 50, 58, 73

CONV Convolutional layer. 50, 51

FC Fully Connected. 50

FCN Fully Convolutional Network. 67

GFDM Gaze fixation density map. 15, 29, 35, 36, 73

GFM Gaze fixation map. 42

HMM Hidden Markov Model. 34

HVS Human visual system. 80, 82, 99

LRN Local Response Normalization. 54

LYLO Les Yeux L'ont. 29

NDD Neuro Degenerative Disease. 29

NN Neural Network. 15, 50, 58

NSS Normalized Scanpath Saliency. 41

POOL Pooling layer. 50, 52

RGB Red Green Blue color space. 24, 79, 82

ROC Receiver Operating Characteristic. 42

SD Standard Definition. 73

SGD Stochastic Gradient Descent. 61

SVM Support vector machine. 65

General introduction

Neurodegenerative diseases with dementia are a real public health problem that increases with the aging of population in developed countries. Indeed, about 860,000 people suffer from dementia of Alzheimer type in subjects older than 65 years in the French population [90]. Tunisian Ministry of Public Health has reported about 20,000 confirmed cases [139]. A timely and non-invasive diagnosis of dementia is essential. Various experiments were performed in order to identify early symptoms of dementia. Oculomotor evaluation is one of them [134].

Before making experiments with Alzheimer patients inclusion, it is important to measure the impact of generated degradations on the visual attention of normal control subjects. Hence, our first goal is to compare visual fixation maps built upon gaze fixations on normal and intentionally degraded video sequences. Furthermore, automatic prediction of visual attention of normal populations has been an intensively researched subject since the last two decades [52]. Here natural video content, but also non-intentionally degraded one due to the coding artifacts and transmission errors [14] were used. It is also interesting to build a predictive model for the normal attention to intentionally degraded content as produced for this study. As the designed degradations are applied to specific areas of interest in the video content, it seems natural to consider machine learning approaches, and in particular Deep Covolutional Neural Networks [130]. Nevertheless, in medical applications visual datasets are usally very small. The reasons are two fold i) limitations in including of patients in medical research experiments, ii) constraints in experiments due to the patients conditions. For the task of saliency prediction the first reason means the impossibility of conducting of psycho-visual experiment on hundreds of patients; and the second reason means that elderly and fragile patients with "NDD" cannot observe visual content during tens of minutes. Hence, the targeted visual content has to be small in volume. In this thesis we try to answer the question how can we use deep learning approaches in such a situation : relatively small amount of measurments on a small database.

Thesis objectives

Three main objectives are involved in this thesis.

First, we wish to better understand the attentional processes that guide the gaze towards particular regions of the visual field.

Second, to model these processes by machine learning tools in view of their approved successes.

And finally to apply this saliency prediction model for testing of patients with neurodegenerative diseases.

The modeling of visual attention with deep convolution networks will allow us to combine the low-level features with the high-level ones for the prediction of regions viewed by a set of subjects when viewing a natural video. Several works and models designed for visual attention with deep learning exist concerned the static image but still quite few who propose to study the videos. We will be interested in this work in the study of dynamic scenes through a database of videos.

We will adapt two complementary approaches. A first one allows to designate the architecture of the deep convolution network ensuring the prediction of salient zones. We define the problem of learning as the bi-class classifictaion problem (salient, Non-salient) with referring to the recordings of the eye movements of subjects viewing natural videos with various contents. The second approach of transfer learning will allow us to propose a model inspired from the already designated architecture to solve the problem of small size of available data sets.

Thesis contributions

In this thesis, we propose to model the saliency in videos by machine learning tools specifically with the deep convolution networks. A designed model allows to insert the residual motion information side by side with the RGB value for each frame of the video. This model, called "ChaboNet4k" classifies input information (residual motion and RGB values) into two classes: salient and Non-salient classes.

The main contributions of the thesis are:

- The proposal of a deep network architecture taking as input : four channels (R, G, B, residual motion), eleven channels (R, G, B, residual motion, and 7 kind of contrasts).
- The study of the effect of data noise on deep networks learning (training and prediction).

- Proposition of a method inspired from wooding method [131] for the generation of a dense saliency map from the probability responses of the proposed deep network model.
- Proposition of a method of transfer learning to solve the problem of few size of available datasets.
- Creation of specific database video for testing of patients with neuro-degenerative diseases.
- Applying transfer learning method for the created specific database.

Thesis outline

In order to be better organized, we have chosen to divide our work into three major parts, each containing two chapters. The first part involves the state-of-the-art on visual saliency prediction and the deep learning for visual saliency prediction. Second part presents the deep CNN designed for prediction of visual saliency in natural video "ChaboNet", and the specific saliecny features used for traning deep CNN. Third part is dedicated for the transfer learning. One chapter was for the proposition of the transfer learning with deep CNN for saliency prediction and the other chapter was for the application of saliency prediction for testing of patients with neuro-degenerative diseases.

Part I

State-of-the-art on visual saliency and deep learning

The goal of this Ph.D thesis is to brought a saliency model that considers the significant content of natural videos. Since the last two decades, saliency prediction in images and digital video is extensively studied by the research community. The current trend in this research topic is the use of deep convolutional networks in order to integrates the semantic aspect.

The outline of this part is as follows: Chapter 1 provides information about the visual saliency prediction. Chapter 2 describes deep convolutional networks for the saliency prediction task.

Chapter 1

Visual Saliency prediction

1.1 Introduction

The oculomotor and particularly rapid eye movements are at the interface of decision-making and motor systems of spatial working memory processes. The study of voluntary saccades has renewed its interest for neurodegenerative diseases "NDD" diagnostics due to the recording simplicity thanks to technical progress and automatic signal analysis. The integration of visual perception of natural scenes in the classification of patients and assessment of disease progression in experimental conditions approaching the ecological situation represents a real scientific challenge. The classification of the degree of disease is based on multiple indicators such as the distribution of the amplitudes of saccades and fixation times, but also on the relationship between the visual fixation maps of patients and normal control subjects. The differences in occulomotor behavior of normal control subjects and subjects with Alzheimer disease, due to the lack of curiosity, with regard to intentionally degraded still images, were reported in [123]. The experiments conducted in the framework of LYLO project "Les Yeux L'Ont" [123]: ocular saccade abnormalities in prodromal Alzheimer's disease", at the University Hospital of Bordeaux (CHU) were also devoted to studies of such phenomena. Such differences can be measured via comparison of visual fixation density maps, GFDM [131] built upon recorded gaze fixations. The goal of the present chapter consists first in understanding the anatomy of human visual system. Second, it gives the state-of-the-art on visual saliency modeling. Finally, this chapter mades an overview on applying saliency prediction for neuro-degenerative disease studies.

1.2 Human visual System

1.2.1 The human eye

In order to study the internal morphology of the eyeball, the following figure 1.1[1]presents a median sagittal section of the eye. Three different tunics are present: the fibrous tunic, the uvea tunic and the nerve tunic. The fibrous or external tunic consists of the opaque sclera (white of the eye) in backward and the transparent cornea toward the front. The uvea tunic consists of three elements iris, ciliary body and choroid. Here the nerve tunic that consists of the retina, is well descriped in order to understand the transduction of the luminous message coming from the outside into nerve signals sent to the brain. Two areas are distinguished from the retina: the visual retina which is defined by the presence of detecting cells. The disappearance of these cells will transform the retina into a simple epithelial seating in the anterior part of the eye which constitutes the blind retina.

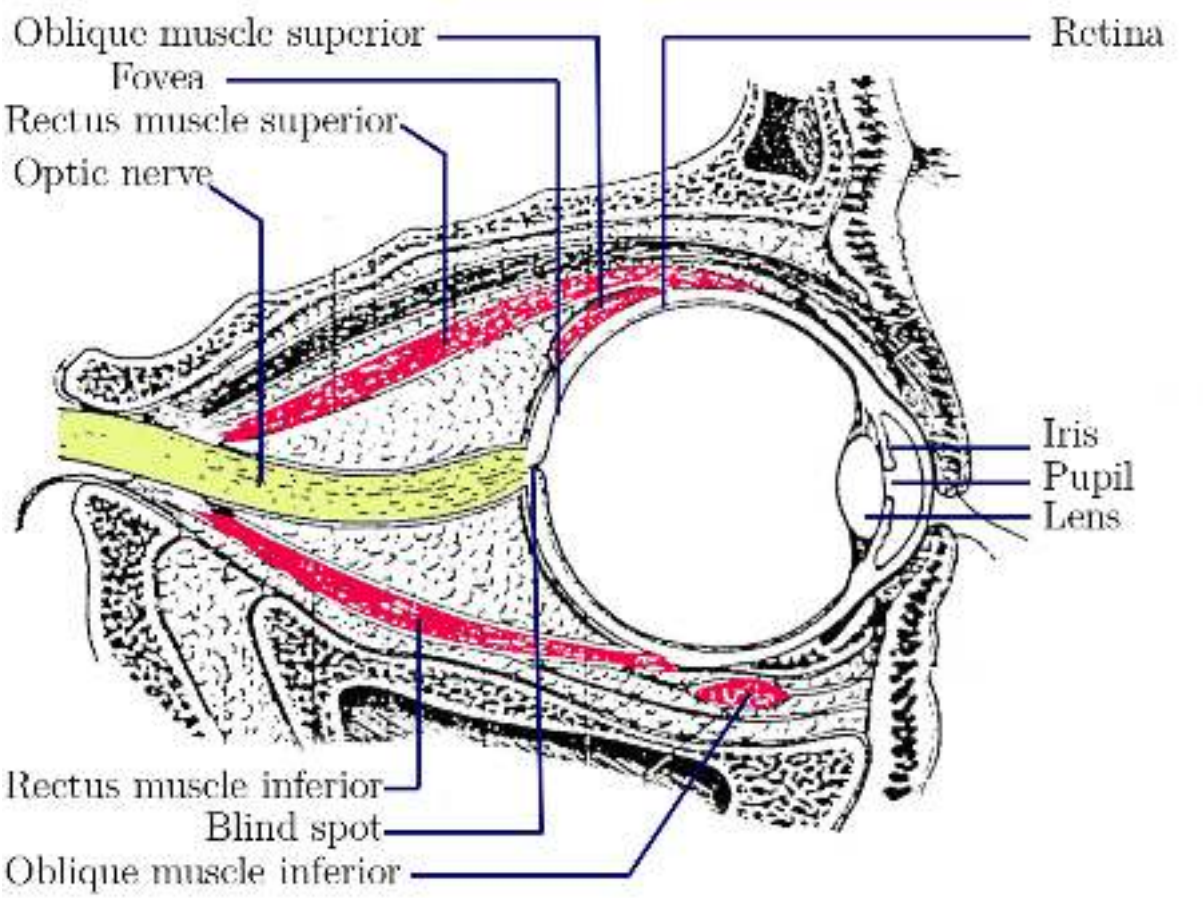


Figure 1.1: Sagittal section of the eye [1]

The central area of the visual retina called macula, presents the daytime and accurate

[1] *https : //theodora.com/anatomy/the_accessory_organs_of_the_eye.html*

viewing area as it can see the colors, shapes and details. The peripheral area specialized in night vision can see just the details. The retina is composed of five layers of neurons (see figure 1.2): photoreceptors, horizontal cells, bipolar cells, amacrine cells and ganglion cells. Optical fibers coming from ganglion cells meet on a disc called the optic papilla which corresponds to the birth of the optic nerve.

Photoreceptors: they constitute the deepest layer of the retina. Two groups of photoreceptors (cones and sticks) are distributed unevenly on the retina. Cones are color sensitive. They intervene in daytime vision. The sticks are involved in the detection of low light intensities and night vision. The photoreceptors which are interconnected in order to smooth the visual information are connected to the bipolar cells and to the horizontal cells.

The horizontal cells which are interconnected in order to smooth the information coming from the photoreceptors, convey information of average luminance to the bipolar cells.

Bipolar cells that connect photoreceptors to a ganglion cell are sensitive to spatial luminance contrast through the center-surround mechanism.

The amarcin cells that laterally share the signal to modulate the response gain of bipolar and ganglion cells are sensitive to temporal contrast and play a role in the detection of motion.

The ganglion cells which constitute the last neuronal layer of the retina transmit the nervous signal in the form of action potentials. Their axons meet to form the optic nerve.

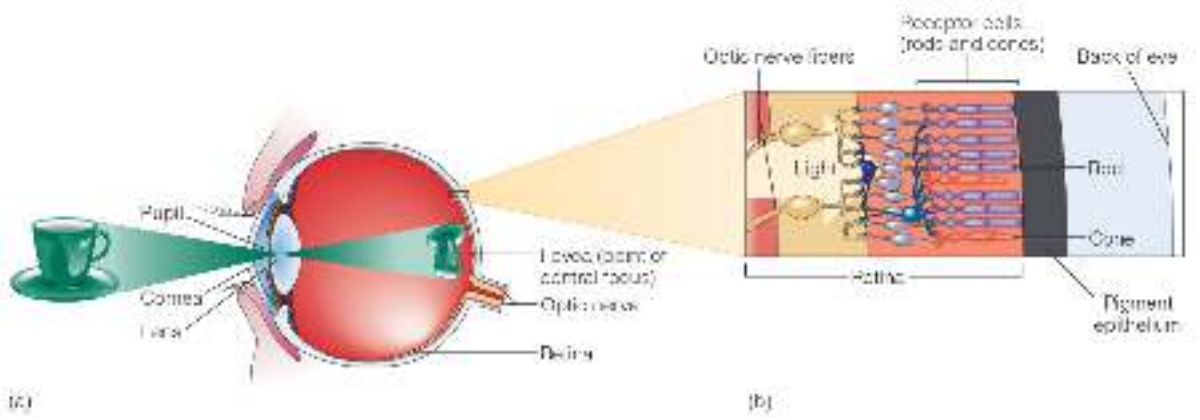


Figure 1.2: An image of the cup is focused on the retina, which lines the back of the eye. The close-up of the retina on the right shows the receptors and other neurons that make up the retina. [36]

1.2.2 Eye movements

Six oculomotor muscles ensure the displacement of the eyeball (see figure 1.3[2]): Four rectus muscles, superior, inferior, lateral and medial; And two oblique, superior and inferior. The superior rectus is an elevator. Its antagonizes, the inferior rectus ensures the depression. The lateral rectus is an abductor which carries the cornea outside. Its antagonizes, the medial rectus is adductor which carries the cornea inside. The anatomical peculiarity of the retina, detailed in above section, pushes the human to move his eyes. The density of photoreceptors on the central area of the fovea (about 5 degree of the visual field) compared to the peripheral zone impacts on the resolution of the visual information. That explain the necessity to move the gaze in order to have the region that we want to analyze in detail in the center of the retina which gives the best visual acuity.

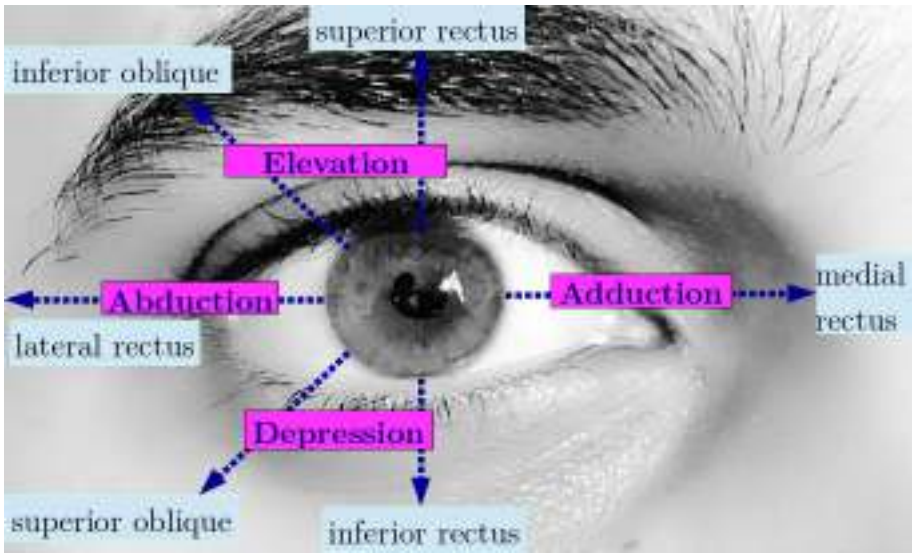


Figure 1.3: Eye motion: Field of action of the oculomotor muscles (right eye[2]).

The saccades, the smooth pursuit and the micro-saccades are the three main eye movements.

- The saccades are a very rapid movements which ensures the setting of the region of interest in the center of the fovea. The movement of the saccades is extremely fast between 30 and 80 ms. Between two saccades the eye stops moving to fix a region for a variable length of time. This period is called a fixation and generally lasts between 250 and 500 ms during this period the visual information is treated.

- smooth pursuit movements allow the tracking of a moving object with a slow speed with a maximum of 100 ms . If the eyes follow the moving object correctly the

[2](Photo credits Wael CHAABOUNI.)

image of the object is stationary on the retina and remains in foveal vision, allowing the visual system to extract more information about the moving object. These movements are continually corrected so as to track the object.

- The micro-saccades are small movements allows the refreshment of the image on the photoreceptors.

Perceiving a motion is executed when something moves across the field of view. Actual motion of an object is called real motion [36]. Apparent motion involves stimuli that are not moving. It is when two stimuli in slightly different locations are alternated with the correct timing just like motion perceives in movies. Induced motion occurs when motion of one object causes a nearby stationary object to appear to move. Motion aftereffects occur after viewing a moving stimulus for 30 to 60 seconds and then viewing a stationary stimulus, which appears to move.

1.2.3 Depth percetion

By controlling the eye axes and the lens focus using the eye muscles, depth estimation is possible with ocolumotor cues. The interaction of this kind of cues ensure the counting of the convergence and accomodation. The angle of convergence of the two eyes and their accommodative states are one source of scaling information.

- Monocular depth cues work with just one eye.
- Binocular depth cues based on the ability to sense the position of eyes and the tension in eye muscles. The difference in the viewpoint of the images received by two eyes creates the cue of binocular disparity.

1.3 Visual saliency modeling

As the quantity of information that reaches the eyes is very high, the processing of the whole information was obstructed. That explain the focusing of the attention only on a part of visual information. This attentional focus towards a particular region of the visual field will lead to move the eyes towards it. For analyse visual perception, attention shift and assessing user interfaces, visual prediction modeling such as eye-tracking technique was used. Here we can speak on two forms or representations of visual prediction that analysing sequences of fixations: dynamic representation called scanpath, and static one called saliency map. Scanpath prensents a sequences of gaze shifts that follow visual attention over an image. While, saliency map is obtained by convolving the fixation

map which represents the spatial coordinates of the set of visual fixation, by an isotropic bi-dimensional gaussian function [67].

Different researchers focused their works to study and to predict scanpaths. Repetitive scanpaths that are made at multiple viewings of the same stimulus, contribute to where people look. A key prediction of scanpath theory [99] is that the top-down recapitulation of scanpaths but also bottom-up guidance might explained it [29].

Liu and al [79] modeled scanpaths based on low-level feature saliency, spatial position, and semantic content [79]. Here, the image was segmented into regions and the proposed model gaze shifts in terms of transition probabilities from one region to another. Transition probabilities between different image regions were calculated through the differences of YUV color values and five scales of Gabor features and eight orientations features. For spatial position and in order to obtain a random walk with steps in an isotropically random direction and a step length subject to a heavy-tailed distribution, steps were modeled with Cauchy distribution. Finally, for extract the semantic content, Hidden Markov Model (HMM) with a Bag-of-Visual-Words descriptor of image regions were used. Next figure 1.4 presents an illustraion of the gaze shifts from Liu and al model [79] .

Figure 1.4: The left and right images show human scanpath segments and corresponding estimates from Liu and al [79] algorithm, respectively, where the correspondences are indicated by matching colors (Ref. [79]).

Recent research work used deep learning for prediction of scanpaths. Here we can cite the work of Assens et al [5] that sampled scanpath by a stochastic approach. The deep network train a model that take a set of image as input and a saliency volumes that are a presentation of spatial and temporal saliency information for images, as output. They have three axes that represent the width and height of the image, and the temporal dimension. Here, they uses the proposed saliency volumes to generate the scanpaths by determining three keys values. First, the the number of fixations of each scanpath, second the duration in seconds for each fixation, were sampled from their probability distributions learned from the training data. And finally, the location of each fixation

point was generated by sampling the time from the corresponding temporal slice.

Simon [116] proposed a model for automatic scanpath generation using a convolutional neural network and long short-term memory modules due to the temporal nature of eye movement data.

In this section, we are more interested by static representation of visual prediction "saliency map". The subjective saliency maps that are built from eye position measurements, the objective saliency maps that are extracted from image or video signal and the comparison metrics between these two kind of saliency maps were well detailed.

1.3.1 Gaze Fixation Density Map (GFDM)

The visual attention map on the group of subjects - the so-called "subjective saliency map" is constructed with the recorded gaze fixations of all subjects in the group. We obtain a map which collects the density of eye positions. Generally, the subjective saliency map S_g , or fixation dense map "GFDM" is obtained by convolving the fixation map by an isotropic bi-dimentional Gaussian function G_σ [67].

$$S_g(X) = [\frac{1}{N_{obs}} \sum_{i=1}^{N_{obs}} (\sum_{m=1}^{M_{fix}} \delta(X - x_{f^{(m)}}))] * G_\sigma(X) \tag{1.1}$$

where

- X is a vector representing the spatial coordinates,
- $x_{f^{(m)}}$ is the spatial coordinates of the m^{th} visual fixation,
- M_{fix} is the number of visual fixation for the i^{th} observer,
- N_{obs} is the number of observers,
- $\delta(.)$ is the Kronecker symbol $\delta(t) = 1$ if $t = 1$, otherwise $\delta(t) = 0$,

An intensive study to densify a fixation coordinates was proposed by Wooding [131]. The method allows the creation of a density map of fixation from a set of views recorded from an oculometer. This method, tested with more than 5000 participants on the digitized images of paintings of the National Gallery, consists of three stages. The first ensures the application of a two-dimensional Gaussian at the center of the eye measurement. And this allows the computing of the partial saliency map for each gaze record. Then the set of partial saliency maps of all subjects are summed in a global saliency map. Finally, the global map was normalized by its maximum value.

For a more describe these three steps, Wooding proposed to fix the Gaussian σ propagation at an angle α of 2°, based on an imitation of the functioning of the fovea of the human eye which covers an area of 1.5° to 2° of the diameter in the center of the retina.

The Gaussian reflects the projection of the fovea on the screen. To ensure this projection, the Gaussian spread σ is defined as follows:

$$\sigma = R \times D \times tan(\alpha) \tag{1.2}$$

with R is the resolution of the screen in pixels per mm and D must be equal to three times the height of the screen ($3H$) according to ITU-R Rec. BT.500-11 [53]. From the equation 1.1 the partial saliency map $S_g^{'}(I,m)$ of the image I for the measurement m of the eye is calculated according to the following equation [19]:

$$S_g^{'}(I,m) = Ae^{-(\frac{(x-x_{0_m})^2}{2\sigma_x^2}+\frac{(y-y_{0_m})^2}{2\sigma_y^2})} \tag{1.3}$$

where $\sigma_x = \sigma_y = \sigma$ and $A = 1$.

Then all partial saliency maps of all subjects are summed in a global saliency map. At the third step, summed up map is normalized by its maximum value, the so-called "saliency peak" in the image. The final GFDM is computed as follow:

$$S_g(I) = \frac{1}{d}\sum_{m=0}^{N_{obs}} S_g^{'}(I,m) \tag{1.4}$$

where $d = \max_{(x,y)\in S_g}(S_g^{'}(I,m))$ is the highest peak and N_{obs} is the total number of subjects.

The following figure 1.5 shows the fixation map of 21 subjects computed with the Wooding's method.

(a) $\#frame42$ (b) $\#frame783$

Figure 1.5: The GFDM saliency map computed during a free task of visualisation of normal sequences by normal subjects.

In next section, we will describe different saliency models that automatically determine regions that attract human gaze on image or video.

1.3.2 Saliency models

A saliency map is a model of neurobiology and psychology that describes how the striking details of the visual environment in the brain, processed as a priority. This model influences pre-attentive (or automatic) exogenous visual stimuli (reflexive, low-level or bottom-up) or endogenous (top-down). Low level factors are described as luminance, orientation and color. Thus, high-level factors may concern cognitive processes, memory, emotional state or task. Therefore, different models of visual attention are designed to clear the salient areas of an image or sequence of images. These models are divided into so-called bottom-up models and top-down models.

Several saliency models [91], [124], [18] have been proposed in various fields of research which are based on the feature integration theory [125]. These research models the so-called "bottom-up" saliency with the theory that suggests the visual characteristics of low-level as luminance, color, orientation and movement to provoke human gaze attraction [30], [31], [48]. The "bottom-up" models have been extensively studied in the literature [10]. They suffer from insufficiency of low-level features in the feature integration theory framework, especially when the scene contains significant content and semantic objects. In this case, the so-called "top-down" attention [104] becomes prevalent, the human subject observes visual content progressively with increasing the time of looking of the visual sequence. Famous examples of top-down attention guidance which showed that eye movements depend on the current taskis is presented by Yarbus in 1967 [133].

Different models of visual attention are designed to clear the salient areas of an image or sequence of images. The most popular and referenced models are detailed in follow.

Bottum-up models

- **Model of Itti and Koch, 1998:** The general idea of the model [52] is summarized by two steps. The first allows the combination of the features of the multi-scale image into a single topographic saliency map. And the second, ensures the selection of the places frequented in decreasing order of saliency thanks to a network of dynamic neurons. The multi-scale analysis depends on a Gaussian filtering step, followed by a subsampling step. Indeed, for the subsampling step, horizontal and vertical image reduction factors range from 1 : 1 to 1 : 256 in eight octaves. These two stages give rise to pyramidal shapes. To summerize, 42 features maps were computed: six for intensity, 12 for color, and 24 for orientation. Applying a normalization step followed by a merge of the maps with the " across-scale addition " operator Itti gets

the final saliency map. A simple browse on the obtained map for the pixel having the highest value, followed by a feedback inhibition mechanism until a defined threshold, performs the identification of the salient areas in decreasing order (see next figure 1.6).

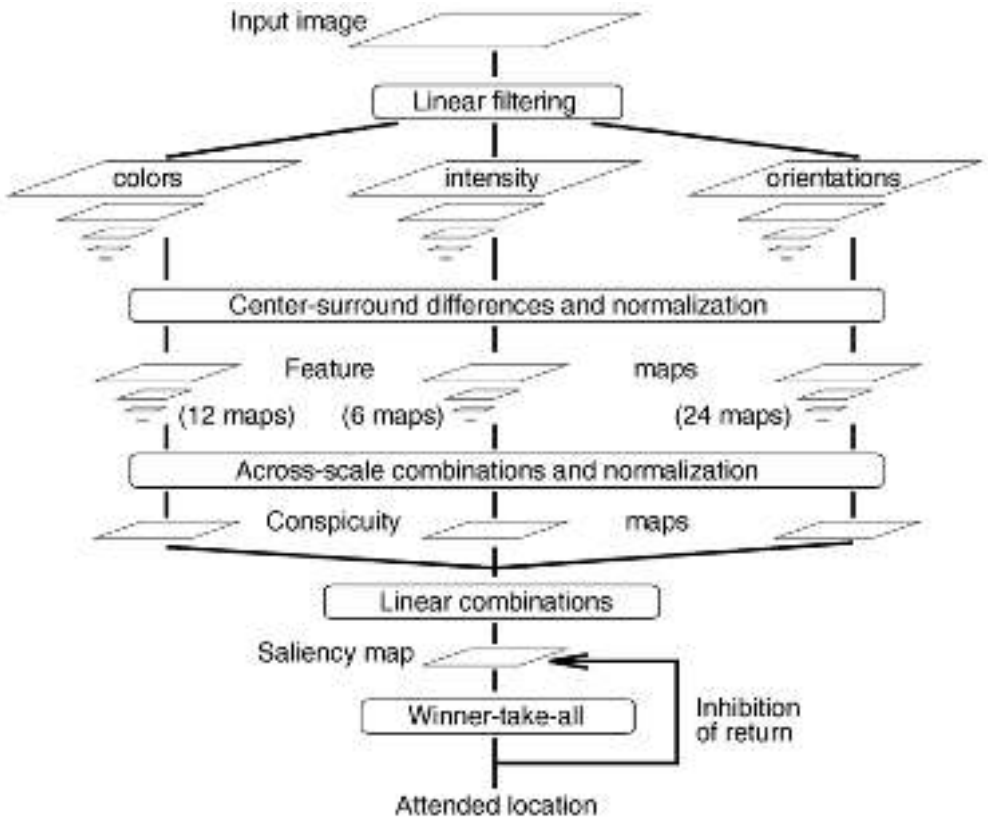


Figure 1.6: Illustration of the architecture of the Itti and Koch model (Ref. [52])

– **Model of Harel GBVS [41] :** Graph-Based Visual Saliency (GBVS) presents a simple, and biologically plausible model that consists of two steps:

i) forming activation maps on certain feature channels that are extracted by linear filtering followed by some elementary nonlinearity. Suppose a given a feature map $Feature_{map} : [n]^2 \longrightarrow R$, to compute an activation map $Activation_{map} : [n]^2 \longrightarrow R$:

$$Activation_{map}(i,j) = -log(p(i,j)) \tag{1.5}$$

where $p(i,j) = Pr\{Feature_{map}(i,j)|neighborhood\}$

ii) normalizing the activation maps in a way which highlights conspicuity and admits combination with other maps. For each node (i,j) and every node (p,q) to wich it

is connected, an edge from (i,j) to (p,q) with weight was introduced:

$$w_2((i,j),(p,q)) = Activation_{map}(p,q).Feature_{map}(i-p, j-q). \quad (1.6)$$

– **Model of Harel signatureSal[47]:** SignatureSal model uses the image signature that is a descriptor of natural scenes. This descriptor can be used to approximate the spatial location of a sparse foreground hidden in a spectrally sparse background. For the problem of figure-ground separation, the spatial support of figure signal is assumed to be sparsely supported in the standard spatial basis. The background is also assumed to be sparsely supported in the basis of the Discrete Cosine Transform. The image signature is defined as

$$ImageSignature(X) = Sign(DCT(X)). \quad (1.7)$$

The figure-ground separation problem is formulated in the framework of sparse signal analysis. The Inverse Discrete Cosine Transform (IDCT) of the image signature concentrates the image energy at the locations of a spatially sparse foreground, relative to a spectrally sparse background.

– **Model of Seo [113]:** presents a bottom-up model that combines static and space-time saliency detection. The space-time saliency detection method does not require explicit motion estimation. First, from a given image or video a local regression kernels was computed and used as features. The use of these kernels ensures obtaining the local structure of images by analyzing the pixel value differences based on estimated gradients. Then, a nonparametric kernel density estimation for such features was used. The saliency map is constructed from a local measure that indicates likelihood of saliency.

– **Model of Marat [86]:** A biologically inspired model (see next figure 1.7) separated a video frame into two signals corresponding to the two main outputs of the retina was proposed by Marat et all [86]. Both signals: spatial information of the visual scene and the motion information, are decomposed into elementary feature maps which are used to form a static saliency map and a dynamic one. These maps are fused into a spatio-temporal saliency map. Three different fusions are used : mean fusion, max fusion and a pixel by pixel multiplicative fusion.

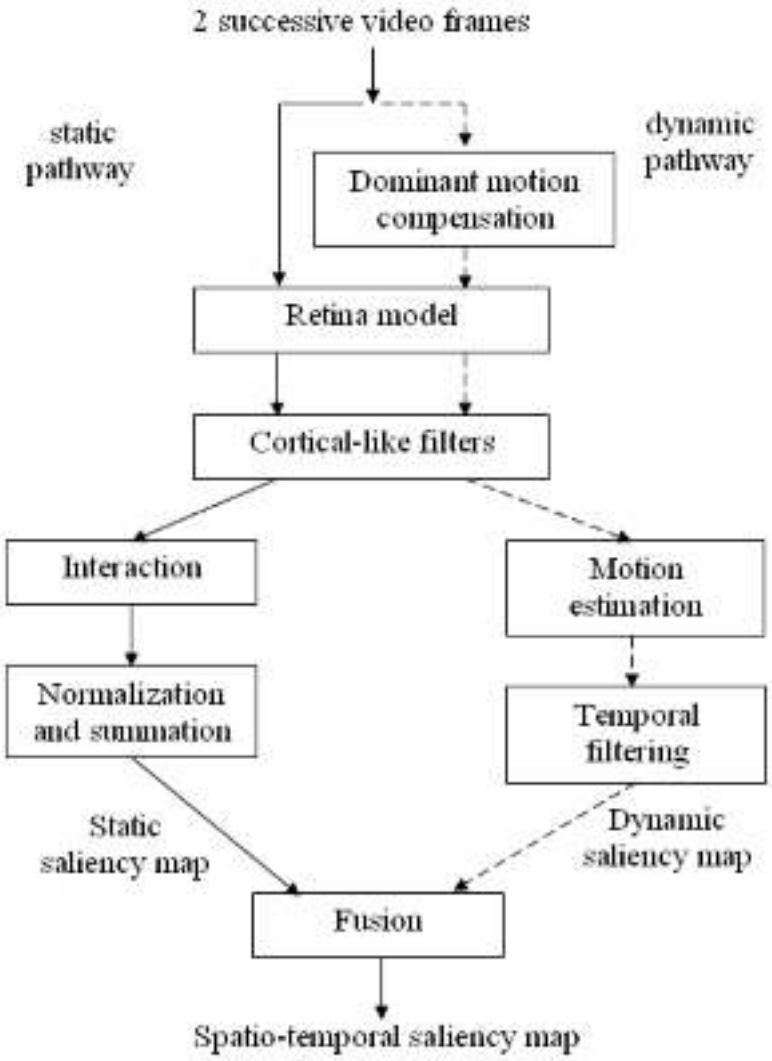


Figure 1.7: Spatio-temporal saliency model for video (Ref. [86]).

Top-down models

Due to the greater difficulty to emulate high-level cognitive process such as scene understanding [49] and task-controlled or objects recognition [19], few researchs were conducted to solve complex vision tasks. Recently, with the high performance of convolutional neural networks on visual tasks, various models has been proposed as a source of top-down attention prediction.

Palazzi [100] aim to predict the driver's focus of attention by answering two major questions: what a person would pay attention to while driving, and which part of the scene around the vehicle is more critical for the task. A multi-path identical deep architecture that integrates visual cues (RGB image), motion by the estimation of optical flow and scene semantics that processes the segmentation prediction on the scene, were proposed.

Each branch of the proposed model is a multiple-input multiple-output architecture designed in the purpose of addressing the strong central bias that occurs in driving gaze data.

Ramanishka et al [108] proposed a top-down saliency approach to expose the region-to-word mapping in modern encoder-decoder networks. This model produces spatial and temporal saliency attention for still images or video. For each word in the sentence, they proposed to compute the saliency value of each item in the input sequence by measuring the decrease in the probability of predicting that word based on observing just that single item.

Murabito et al [94] presented a SalClassNet approach based on CNN framework consisting of two networks jointly trained. the first network "CNN saliency detector" generates a top-down saliency maps from input images that consist of eye-gaze data recorded. And the second "CNN classifier" ensures exploiting the computed saliency maps for visual classification.

1.3.3 Comparison metrics of saliency maps

In the literature, different evaluation metrics were used to determine the likelihood ratio between the saliency maps and the points recording the eye movements. Four metrics allow a simple interpretation of the results: the Pearson and Spearman correlation coefficients used in various domains to judge the similarity of two distributions, the area under the ROC curve allowing the evaluation of the quality of a prediction, and the NSS "Normalized scanpath saliency" which is defined in the studies of visual attention to compare the salient areas determined by a model with the areas observed by the subjects.

- **Normalized scanpath saliency** is a Z-score that express the divergence between saliency map and human visual attention. The aim is to measure the value of the saliency in the fixation zones along the entire length of the gaze path. After normalization of the saliency map S_M in order to have a zero average and a standard deviation equal to one, the NSS value is calculated on a small centered neighborhood for each fixing location [85]. Due to pre-normalization of the saliency map, a positive value of NSS suggest a greater correspondence than expected by chance between the fixation areas and the predicted salient points; a null value indicates the absence of this correspondence, while negative value indicate an anti-correspondence between the fixation points and the salient points. In conclusion, the higher the positive NSS

value, the more the fixed points are salient. NSS is written as follow:

$$NSS = \frac{\overline{S_g \times S_M} - \overline{S_M}}{\sigma(S_M)} \tag{1.8}$$

where $\overline{S_M}$ is the mean of S_M and σ presents its standard deviation.

– **Pearson Correlation Coeficient** is a metric that measures the force and direction of a linear relationship between two saliency maps. The aim is to calculate the intensity of the connection between the saliency map S_M and the gaze fixation map GFM S_g. This intensity reflects the degree of similarity between the two maps. The calculation of the standard deviation of each saliency maps and the covariance between these two maps makes it possible to determine the PCC value. The coefficient PCC is bounded between $[-1\ 1]$. The closer the PCC value is to the upper bound (1), the more the areas viewed correspond to areas of strong saliency. A value of zero indicates the absence of correspondence between the saliency and the eye positions, ie the absence of linear relationship between the two maps. Whereas the negative values (PCC tends to -1) indicate the correspondence of the observed zones with low saliency zones. The following equation calculate the PCC value.

$$PCC(S_M, S_g) = \frac{cov(S_M, S_g)}{\sigma_{S_M}\sigma_{S_g}} \tag{1.9}$$

where, $\text{cov}(S_M, S_g)$ is the covariance between S_M and S_g; σ_{S_M}, σ_{S_g} represent the standard deviation of maps S_M and S_g respectively.

– **Area under the ROC Curve** is a metric that measures the accuracy of a system that categorizes entities into two distinct groups based on their characteristics. The pixels of the image can belong either to the category of pixels viewed by subjects or to the category of pixels that have not been viewed by any subject. The curve is obtained from plotting of the points having as abscissa the rate of false positives and as ordered the rate of true positive. The rate of true positives $TVP = \frac{tp}{(tp+fn)}$ shows the number of pixels fixed by the subjects and having a saliency value greater than the threshold, divided by the total number of pixels fixed. The false positive rate $TFP = \frac{fp}{(fp+tn)}$ collects the number of pixels with a saliency value higher than the threshold but which have not been fixed and divides it by the number of pixels not fixed. The larger of the area, the more the curve deviates from the random classifier line (area 0.5) and approximate the ideal classifier (area of 1.00). A value close to 1 of Auc indicates a correspondence between the saliency map and the gaze fixations.

While a value close to 0.5 presents a random generation of the saliency zones by the model. And then the objective and subjective maps are very dissimilar. The following algorithm 1 defines the instructions for calculating the AUC.

Algorithm 1 compute_AUC

Require: $\{S_g\}$: map (pixels vector) of gaze fixation
$\{S_M\}$: objective saliency map
$\{subj\_threshold\}$: threshold of S_g

Ensure: $\{auc\_value\}$: value of AUC metric.

$m\_thresholdTab[nbr]$: nbr thresholds uniformly distributed between the min and max of the map S_M

for for each value *count* of the table $m\_thresholdTab[nbr]$ **do**
  for for each pixel i of the frame **do**

    if $(S_g[i] \geq subj\_threshold)$ **then**

      if $(S_M[i] \geq m\_thresholdTab[count])$ **then**
        $++tp$: increase of the number of true positive
      else
        $++fn$: increase of the number of false negative
      end if
    else

      if $(S_g[i] \geq m\_thresholdTab[count])$ **then**
        $++fp$: increase of the number of false positive
      else
        $++tn$: increase of the number of true negative
      end if
    end if
  end for
  calculation of the True Positive Rate : $TVP[count] = \frac{tp}{(tp+fn)}$
  calculation of the false Positive Rate : $TFP[count] = \frac{fp}{(fp+tn)}$
end for
for for each value *count* of the table $m\_thresholdTab[nbr]$ **do**
  $auc\_value+ = (\frac{(TVP[count-1]+TVP[count])}{2}) \times (TFP[count] - TFP[count-1])$
end for
return *auc_value*

– **Spearman's Ranc-Order Correlation** is a metric that measures the correlation between the ranks of the values taken from the two variables rather than the exact values. Since the PCC and the AUC area should vary jointly to some degree even if they have different objectives [26], the SROC coefficient was computed to identify the degree of interaction between these two metrics. To determine this coefficient, one rank is assigned for each PCC value and AUC calculated from the saliency maps of each frame of the video sequence. The calculation of the Pearson correlation coefficient between the ranks of the PCC and AUC values of each frame allows us to obtain the value of the coefficient SROC. The sign of the Spearman correlation indicates the direction of binding between PCC and AUC. If AUC tends to increase when PCC increases, Spearman's correlation coefficient is positive. If AUC tends to decrease when PCC increases, Spearman's correlation coefficient is negative. A Spearman correlation of zero indicates that there is no tendency for AUC to increase or decrease when PCC increases. The Spearman correlation increases in magnitude as PCC and AUC approximate being perfect monotone functions. For calculating the SROC metric where $Rank_{PCC}$ and $Rank_{AUC}$ present the ranks of the scores PCC and AUC respectively:

$$SROC(Rank_{PCC}, Rank_{AUC}) = \frac{cov(Rank_{PCC}, Rank_{AUC})}{\sigma_{Rank_{PCC}} \sigma_{Rank_{AUC}}} \tag{1.10}$$

1.4 Saliency prediction for NDD studies

Neurodegenerative diseases mainly affecting neurons, cause damage to the nervous system (brain and spinal cord). And since neurons are not renewed, damage or death of a neuron can never be replaced. For this, preventive treatment is essential in order to fight against these diseases. The clinical diagnosis of these diseases, is based on finding specific symptoms of disease syndromes. In addition to the cognitive disturbances, slowness, stiffness and tremors that are the main symptoms of neurodegenerative diseases, several studies have to prove that an oculomotor evaluation makes it possible to diagnose these diseases [35] [83]. Oculomotricity and in particular rapid eye movements are at the interface of decision-making engine systems and spatial work memory processes. The study of voluntary saccades benefits from a renewed interest in neurodegenerative pathologies due to the simplicity of recording thanks to technical advances and automatic analysis of the signal. The integration of the visual perception modeling of natural scenes into patient classification and the quantification of disease progress under experimental conditions approaching the ecological situation represents a real scientific challenge.

The classification and quantification of the degree of disease in patients is based on multiple indicators, such as the distribution of saccade amplitudes and duration of fixation, and also on the relationship between the visual fixation maps of patients and control subjects. Delays in the oculomotor function of patients with neurodegenerative diseases must be characterized by a time lag in the fixation maps. Other differences are hypothetically expected. Neurodegenerative diseases show eye movement disorders. Indeed, and contrary to their slowed-down movement, people with Parkinson's disease produce automatic rapid movements of the eye to sensory stimuli and show an impairment of the ability to generate voluntary eye movements in cognitive tasks. The study of [20] has shown that participants with Parkinson's have deficits in their ability to inhibit automatic saccades (more express, more errors in direction ...). Thus they take longer time for volitious jerks (anti-saccade task). And regarding the processes of spatial memory work, Parkinsonians show deficits in moving their eyes to goals called in the right order. To determine the mental state of patients with neuro-degenerative disease and the evolution of the disease, different experiments have been put in place. In the next section, we will describe the experiment carried out by [126] allowing the classification of clinical populations from the natural vision ocular movements and the [4] study which allows to examine the error rates and visual exploration strategies of Parkinson's patients.

1.4.1 Experiment of Tseng,2013 [126]

To extract the essential characteristics that differentiate patients from control subjects, Tseng [126] used automatic learning in a workflow inspired by microarrays analysis. Indeed, this experiment involved two configurations of eye tracking (one for children and another for young adult subjects) but it is still identical in field of view to the stimuli. Participants who sit in front of the screen, watch ten videos of one minute each. The right eye of the observers was measured at 500 Hz.

To create the learning model, Tseng [126] used ten saliency maps: nine were extracted from the various low-level visual features, and one top-down map was generated by the instantaneous viewing positions of 19 of young adults. The nine saliency maps are created from the itti saliency model [52]. Tseng [126] used the Itti model to identify visually highlight regions that can attract the attention on natural videos. As a result, all ten saliency maps provide information that controls attention from top-down in addition to low-level features. The following figure 1.8 summarizes the evaluation of the deployment of the attention proposed by Tseng [126]. In fact the movements of the eyes of the observers are recorded (red curve) during the free viewing of videos of natural scenes. And, the implementation of the architecture of Itti's model was extended (C, color , I, intensity,

O, orientation, F, flicker, M, motion, J, junction of line).

Based on previous studies of high prevalence neurological disorders involve oculomotor and attention deficit dysfunctions, Tseng [126] extracted a large number of characteristics (224) from the eye movement records, and then based on its characteristics, they constructed a classifier to differentiate patients from healthy subjects.

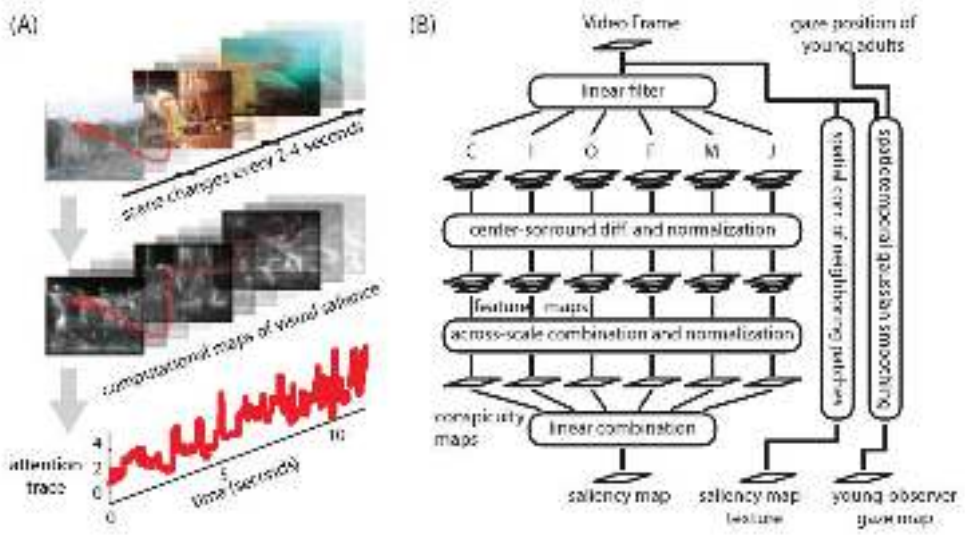


Figure 1.8: Evaluation of the deployment of visual attention: 'A' presents the extracts of the traces of the attention. 'B', it summarizes the extended architecture of Itti's model. (Ref. [127])

1.4.2 Experiment of Archibald,2013 [4]

Abnormal eye movements, such as the depreciation of rapid eye movements (saccades) and the interspersed fixings appeared under the influence of cortical and subcortical networks often targeted by neurodegeneration seen in Parkinson's disease. From this marker of cognitive decline, the study of Archibald [4] examines the error rates and visual exploration strategies of Parkinson's with and without cognitive impairment. Here, the creation of a predictive model of the fixation duration from a data analysis of the tasks, makes it possible to predict both cognitive disorders and severity of the disease.

The stimuli used in this study are presented in five blocks (Figure 1.9): a task corresponding to the angle, a task corresponding to the clock and the reverse clock, a shape position task and finally a task of overlapping figures. Each block consists of 16 test images and is arranged in such a way that a stimulus is presented at the center and four comparators are arranged just below it. Stimuli were thrown on a 20-inch screen at a distance of 80 cm to participants in a dimly lit room.

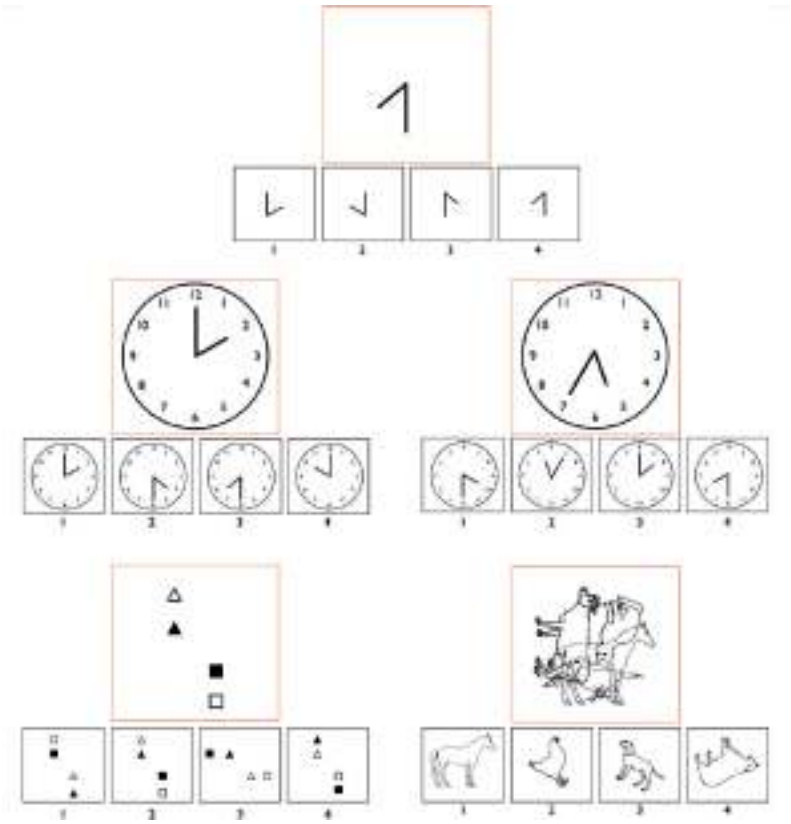


Figure 1.9: Test used in the eye-tracking battery. (Ref.[4])

The first hypothesis of the Archibald [4] study concerning the exploration strategy, as defined by the time of the first correct fixation, the number of central passages and the number of passages, shows that patients with dementia differ in all effectiveness of the exploration strategy compared to cognitively normal Parkinson's patients and the healthy subjects. Second, this study shows that there is a small, but significant, difference in fixation time. The cognitively normal subjects in the Parkinson's disease group made fixations always slower than the control subjects, of the order of 18 ms. This prolongation of the duration of fixation was more pronounced in subjects with Parkinson's disease with dementia.

1.5 Conclusion

In this chapter, a state-of-the-art of visual saliency prediction was provided after detailing the anatomy of human eyes . We briefly presented the biology of the human visual system, especially the retina and ocular movements. Hence, in order to perceive the world, human attention was focused on small regions in order to receive more detail. These regions are selected according to attentional processes "bottum-up" which depends on low level factors and "top-down" which may concern cognitive processes, memory, emotional state

or task. The visual attention precedes the displacement of the human gaze through various ocular movements towards the region on which the visual attention is directed.

We discussed the commonly saliency prediction model and described the ones we considered most relevant for the rest of our work. These models are inspired by the feature integration theory and the recent deep convolutional networks that ensures the combination of "top-down" and "bottum-up" visual stimuli.

Then we introduced how these models and what kind of real-life applications can emerge from it, precisely, for neuro-degenerative diseases studies. Hence, rapid eye movements are the interface of spatial work memory processes.

In next chapter, we explain deep convolutional networks used for predict visual attention.

Chapter 2

Deep learning for visual saliency prediction

2.1 Introduction

Machine Learning is a set of techniques used to achieve, automatically, a task by learning from a training data set. There is a plethora of methods based on different mathematical fundamentals. Neural networks were intended to model learning and pattern recognition done by physiological neurons. This was first introduced by Hebb (1949) who modeled synapses by weighted links from the outputs of nodes to the inputs of other nodes. Rosenblatt (1958) continued the Hebb model and investigated how the links between neurons could be developed, in particular, he defined the basic mathematical model for neural networks (NN for short). His basic unit was called the perceptron, which when it receives a signal, would either respond or not, depending on whether a function exceeded a threshold. Figure 2.1 presents a formal neurone. It receives input signals $(x_1, x_2, \cdots, x_p)$, and applies an activation function f to a linear combination of the signals. This combination is determined by a vector of weights $w_1, w_2, \cdots, w_p$ and a bias b_0. More formally, the output neurone value y defined as follows:

$$y = f\left(b_0 + \sum_{i=1}^{p} w_i x_i\right). \tag{2.1}$$

A neural network is then a network whose nodes are formal neurones, and to define a neural network, one needs to design its architecture (the number of hidden layers and the number of nodes per layer, etc) as well as estimation of parameters once the network is fixed. Figure 2.2 gives an example of such a network.

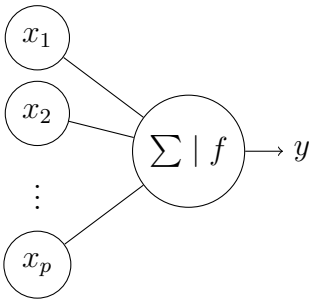


Figure 2.1: A formal neurone.

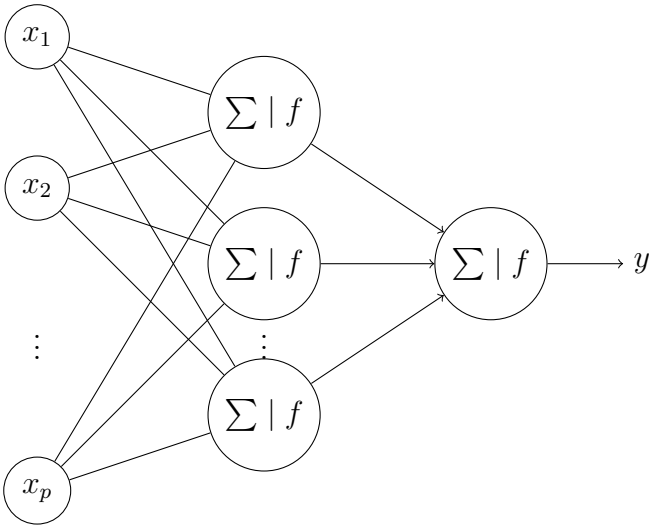


Figure 2.2: An example of a NN. Data X is fed into the first (and here only) hidden layer. Each node in the hidden layer is the composition of a sigmoid function with an affine function of X. The outputs from hidden layer are combined linearly to give the output y.

This chapter consists first in understanding the deep convolutional neural network. Here, the set of steps constituting the design of a convolutional neural network are described: the different commun layers and a state of the art of deep architecture for specific tasks was conducted. Then, the different loss functions and optimization methods were explored. Section 2.4 describes the problem of training Deep CNNs when processing a noisy traning dataset. Section 2.5 presents transfer learning. Finally, this chapter mades an overview on saliency prediction by machine learning.

2.2 Deep Convolutional Neural Networks

Deep learning is a branch of machine learning introduced in 1980s. Nevertheless, its emergence started really by the computational power of the 2000s. It is a machine learning process structured on a so-called convolutional neural network (CNN). A CNN is com-

posed of several stacked layers of different types: convolutional layers (CONV), pooling (POOL) layers, non-linearity layers such as ReLu layers or sigmoid layers, and (generally the last layer) fully connected layers (FC). Figure 2.3 gives an example of an architecture of a CNN.

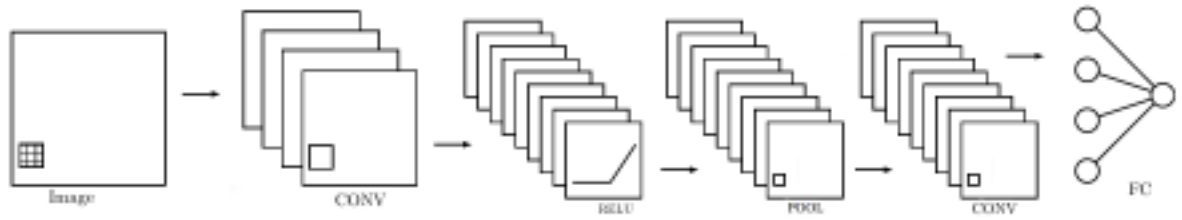


Figure 2.3: An example of a CNN.

2.2.1 Commun layers

Convolutional layers (CONV)

In order to extract the most important information for further analysis or exploitation of image patches, the convolution with a fixed number of filters is needed. It is necessary to determine the size of the convolution kernel to be applied to the input image in order to highlight its areas. Two stages are conceptually necessary to create a convolutional layer. The first refers to the convolution of the input image with linear filters. The second consists in adding a bias term.

Generally, the equation of convolution can be written as (2.2):

$$X_j^l = \sum_{k \in \Omega_j} X_k^{l-1} \odot W_k^l + b_j^l \tag{2.2}$$

with Ω_j - is the kernel support, i.e. the receptive field of j-th neuron;

l- is the netwok layer;

X_j^l - is the input of j-th neuron at layer l, that is feature-map vector;

W_k^l - is the wieght of k-th neuron in the receptive field Ω_j;

b_j^l - is the bias of j-th neuron at the layer l.

$\odot$ is Hadamard product which is a coordinate-wise operation.

In practice, each Conv layer is defined by four parameters: the number of filters K, the spatial extend or the kernel size F, the stride between each region $\tilde{S}$ and finally the amount of zero padding $\tilde{P}$. The Conv layer accepts as input a volume of a size $W1 \times H1 \times D1$ where $W1$, $H1$ and $D1$ present the width, the height and the channels

number respectively that correspond to the input blob (in the first conv layer the blob is the input image). In order to define the four parameters of the Conv layer, some equation should be respected to produces the output volume of size $W2 \times H2 \times D2$:

$$W2 = (W1 - F + 2\tilde{S})/\tilde{P} + 1 \tag{2.3}$$

$$H2 = (H1 - F + 2\tilde{S})/\tilde{P} + 1 \tag{2.4}$$

$$D2 = K \tag{2.5}$$

Pooling layers (POOL)

Pooling reduces the computational complexity for the upper layers and summarizes the outputs of neighboring groups of neurons from the same kernel map. It reduces the size of each input feature map by the acquisition of a value for each receptive field of neurons of the next layer. Different function could be used in the pooling operation such as average or maximum. With the fallen out of average pooling, recent deep networks used max-pooling, see equation (2.6):

$$h_j^n(x, y) = \max_{\bar{x},\bar{y}\in\mathcal{N}} h_j^{n-1}(\bar{x}, \bar{y}) \tag{2.6}$$

Here $\mathcal{N}$ denotes the neighborhood of (x,y).

In practice, each POOL layer is defined by two parameters: the spatial extend or the kernel size F, the stride between each region $\tilde{S}$. Commonly, these parameters are defined as $F = 2$ and $\tilde{S} = 2$; but we can also used the overlapping pooling with $F = 3$ and $\tilde{S} = 2$. The Pool layer accepts as input a volume of a size $W1 \times H1 \times D1$ and produces the output volume of size $W2 \times H2 \times D2$ where:

$$W2 = (W1 - F)/\tilde{S} + 1 \tag{2.7}$$

$$H2 = (H1 - F)/\tilde{S} + 1 \tag{2.8}$$

$$D2 = D1 \tag{2.9}$$

Activation layers

Acitvation layers used a non-linearity function that takes a single number and performs a certain fixed mathematical operation on it. Here, we will describe several activation functions :

ReLu layers

The Rectified Linear Unit (ReLu for short) has become very popular in the last few years. It computes the function $f(x) = \max(0, x)$ (see figure 2.4). Thus, the activation is thresholded at zero. It was found to accelerate the convergence of a very popular parameter optimization method, stochastic gradient descent, compared to the sigmoid function.

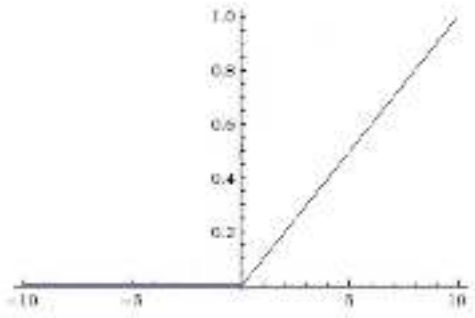


Figure 2.4: Rectified Linear Unit (ReLU) activation function

A first variation of ReLu layer was available to resolve the "dying ReLU" problem : Parameterized Rectified Linear Unit [42] where a non-linearity function is applied $f(x_i) = \max(0, x_i) + a_i \min(0, x_i)$ where a_i is a small constant. The differences from ReLULayer are 1) negative slopes (of 0.01, or so) are learnable though backprop and 2) negative slopes can vary across channels.

The second variation of Relu layer generalizes the ReLU and its first variation to be written as $f(W^T x + b)$ [38]. Here, the dot product between the weights W^T and the data x presents a non-linearity function.

Sigmoid layers The sigmoid non-linearity takes a real-valued number and "squashes" it into range between 0 and 1 (see figure 2.5). It has the mathematical form $\sigma(x) = \frac{1}{1+e^{-x}}$. It was well used in neural network since its nice interpretation (large negative numbers become 0 and large positive numbers become 1). With convolutional network, the sigmoid function saturates at either of 0 or 1 and then it kills gradients.

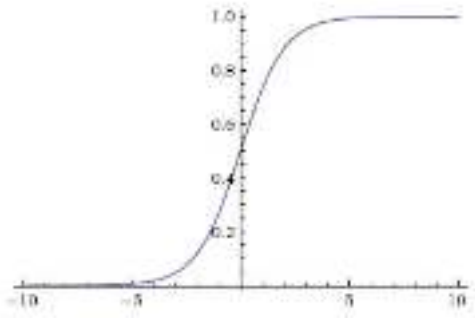


Figure 2.5: Sigmoid activation function

TanH layers

The tanh squashes a real-valued number to the range $[-1, 1]$ (see figure 2.6). The tanh neuron with the non-linearity function $f(x) = 2\sigma(2x) - 1$ is a scaled sigmoid neuron. Here, its activations saturate and its output is zero-centered.

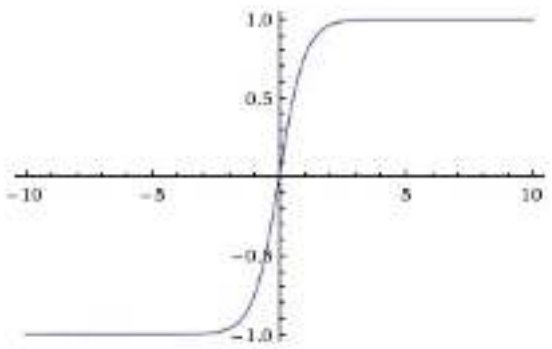


Figure 2.6: TanH activation function

Local response normalization layers (LRN and ReLu)

A local Response Normalization (LRN) layer normalizes values of feature maps which are calculated through the neurons having unbounded (due to ReLU) activations to detect the high-frequency characteristics with a high response of the neuron, and to scale down answers that are uniformly greater in a local area. The output computation is presented in equation (2.10):

$$\psi(Z(x,y)) = \frac{Z(x,y)}{(1+\frac{\alpha}{N^2}\sum_{x'=\max(0,x-[N/2])}^{\min(S,x-[N/2]+N)}\sum_{y'=\max(0,y-[N/2])}^{\min(S,y-[N/2]+N)}(Z(x',y'))^2)^{\beta}} \tag{2.10}$$

Here $Z(x, y)$ represents the value of the feature map after ReLU operation at (x, y) coordinates and the sums are taken in the neighbourhood of (x, y) of size $N \times N$, α and β regulate normalization strength. Normalization is also a coordinate-wise opeartion.

Fully-connected layer

Neurons in a fully connected layer have full connections to all activations in the previous layer, as seen in regular Neural Networks (see figure 2.2). Their activations can hence be computed with a matrix multiplication followed by a bias offset. As Convolution layer, the Fully-connected layer compute dot products. The only difference between these two layers is that the neurons in the Convolution layer are connected only to a local region in the input.

2.2.2 Deep CNN architecture for specific tasks

As detailed before, the main layers of a deep CNN are the convolutional layer which will compute the output of neurons by a dot product between their weights and the connected local regions in the input. Generally, the convolutional layer is followed by an elementwise activation function which leaves the size of the volume unchanged. Using deep networks for classification in image processing, ensures the transformation of the input image pixels considered as neurons to finally yield a single output that presents the label of the input image. Indeed, the input values go through the network, undergoing subsampling, non linear transformation and linear combination as they pass through the layers to finally classify the image. Each neuron in the CNN can be seen as a feature extractor, by applying a filter to the image. The inputs of the intermediate layers are the result of a combination of filters from the layer above. This means that neurones of the first layers extract a "simple" features and those in deep layers are used to extract more complex features.

Historically, the first hypothetical neural machine was illustrated by the perceptron of Rosenblatt [111]. It presents the analogy to biological systems. In next figure 2.7, the sensory units of retina response with an all-or-nothing to the stimilus intensity. These impulses are transmetted to a set of association cells in a projection area. Here, each cells receive a number of connections from the sensory points. Connections between the projection area and the association area A_{II} are random. The R_1,R_2, . . . ,R_n cells response like the units of association area. Here, a feedback connections between cells response and the association area are used.

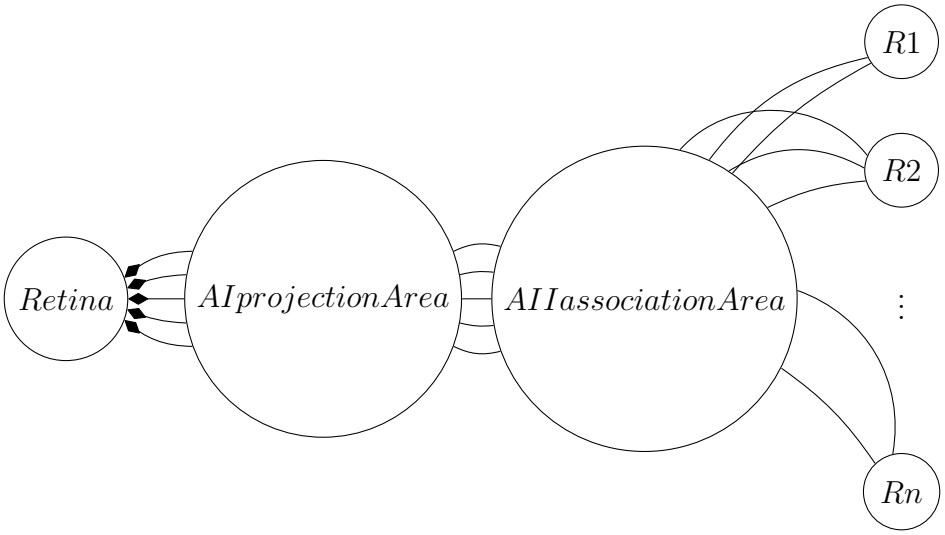


Figure 2.7: Organization of the perceptron of Rosenblatt [111] : localized connection betwen the retina and AI projection area; random connection otherwise.

The authors of [32] designed a fast and reliable face detection system using convolutional neural networks, to detect face patterns of variable size and appearance, that are rotated up to ±20 degrees in image plane and turned up to ±60 degrees, in complex real world images. The proposed CNN consisted of a two convolutional layers which ensure the feature extraction, each one is followed by a sapsamling layer which reduce of dimensionality (average Pooling) (see figure 2.8). Fully connected layers $N1$ and $N2$ contain simple sigmoid neurons in order to perform classification of "face" or "no-face" problem.

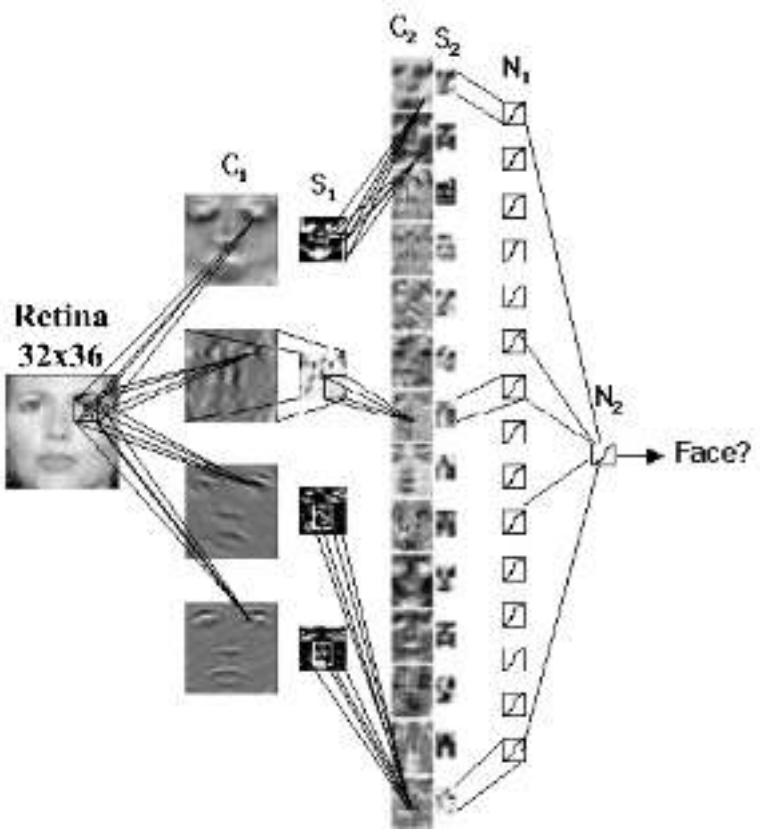


Figure 2.8: Architecture of face detection network. (Ref. [32])

The first successful applications of Convolutional Networks was proposed by LeCun in [69]. Here, a convolutional neural network LeNet was specifically designed for on-line handwriting recognition. As illustrated in figure 2.9, the deep network is constructed with 7 layers. The lower-layers are composed to alternating convolution and max-pooling layers. The upper-layers however are fully-connected and correspond to a traditional MLP with a logistic regression (see figure 2.2). The proposed convolutional neural network eliminates the need for hand-crafted features extractors and reduce the need of hand-crafted heuristics and manual parameter tuning in document recognition systems.

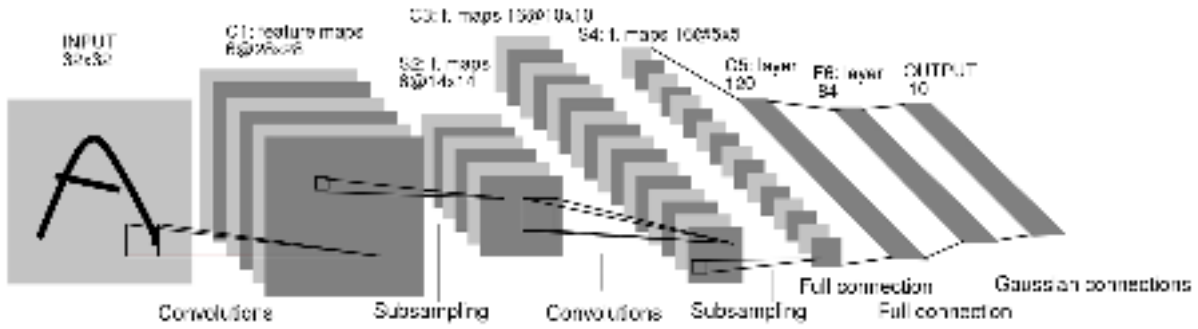


Figure 2.9: Architecture of LeNet network. (Ref. [69])

The popular deep convolutional neural network proposed by Krizhevsky [62], used eight layers with weights to classify the 1.2 million high-resolution images [112] into a 1000 different classes. This depth architecture achieved a record-breaking results using purely supervised learning. The first five layers are convolutional and the remaining three layers are fully connected. The output of the last fully-connected layer is fed to a 1000-way softmax which produces a distribution over the 1000 class labels. The network used to maximize the multinomial logistic regression. The neurons in the fullyconnected layers are connected to all neurons in the previous layer. Response-normalization layers follow the first and second convolutional layers. Max-pooling layers. The ReLU non-linearity is applied to the output of every convolutional and fully-connected layer

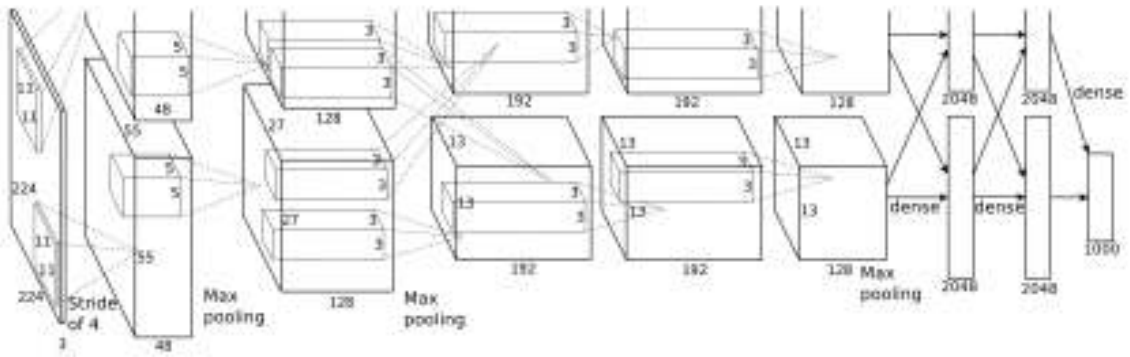


Figure 2.10: Architecture of AlexNet network for object recognition. (Ref. [62])

Today with the different frameworks available for deep networks training, such as tensor flow [1], torch [21], Caffe [54], Theano [9], ... an explosion of network architecture for different tasks has emerged ZF-Net [136] , GoogLeNet [121], VGGNet [118] and ResNet [43].

2.3 Loss Functions and Optimization Methods

A neural network be it a fully connected NN or a CNN is a supervised machine learning model. It learns a prediction function from a training set[129]. Each sample from this set can be modeled by a vector which describes the observation and its corresponding response. The learning model aims to construct a function which can be used to predict the responses for new observations while committing a prediction error as lowest as possible.

More formally, a sample i from the training set is denoted $(x_1^i, x_2^i, \cdots, x_n^i, y^i)$ and the response of the model is denoted $\hat{y}^i$.

2.3.1 Loss functions

There are many functions used to measure prediction errors. They are called *loss functions*. A loss function somehow quantifies the deviation of the output of the model from the correct response. We are speaking here about "empirical loss" functions [129], that is the error computed on all available ground truth training data. Here we will shortly present one of them.

One-hot encoding

Back to the training set, the known response of each observation is encoded in a one-hot labels vector. More formally, given an observation $(x_1^i, x_2^i, \cdots, x_n^i, y^i)$, we introduce a binary vector $L^i = (L_1^i, L_2^i, \cdots, L_k^i)$ such that if $y^i = c_j$ then $L_j^i = 1$ and $\forall m \neq j$, $L_m^i = 0$. This is the function which ensures a "hard" coding of class labels.

Softmax

Given a vector $Y = (y_1, y_2, \cdots, y_k)$ with positive real-valued coordinates, the softmax function aims to transform the values of Y to a vector $S = (p_1, p_2, \cdots, p_k)$ of real values in the range $(0, 1)$ that sums to 1. More precisely, it is defined for each $i \in \{1, 2, \cdots, k\}$ by:

$$p_i = \frac{e^{y_i}}{\sum_{j=1}^{k} e^{y_j}}. \tag{2.11}$$

The softmax function is used in the last layer of multi-layer neural networks which are trained under a cross-entropy (we will define this function in next paragraphs) regime. When used for image recognition, the softmax computes the estimated probabilities, for each input data, of being in a class from a given taxonomy.

Cross-Entropy

The cross-entropy loss function is expressed in terms of the result of the softmax and the one-hot encoding. It is defined as follows:

$$O(S, L) = -\sum_{i=1}^{k} L_i \log(p_i). \tag{2.12}$$

The definition of one-hot encoding and the equation (2.12) means that only the output of the classifier corresponding to the correct class label is included in the cost.

Average Cross Entropy

To deal with the cross-entropy of all the training set, we introduce the average cross-entropy. This is simply the average value, over all the set, of the cross-entropy introduced in equation (2.12):

$$\mathscr{L} = \frac{1}{N} \sum_{i=1}^{N} O(S^i, L^i). \tag{2.13}$$

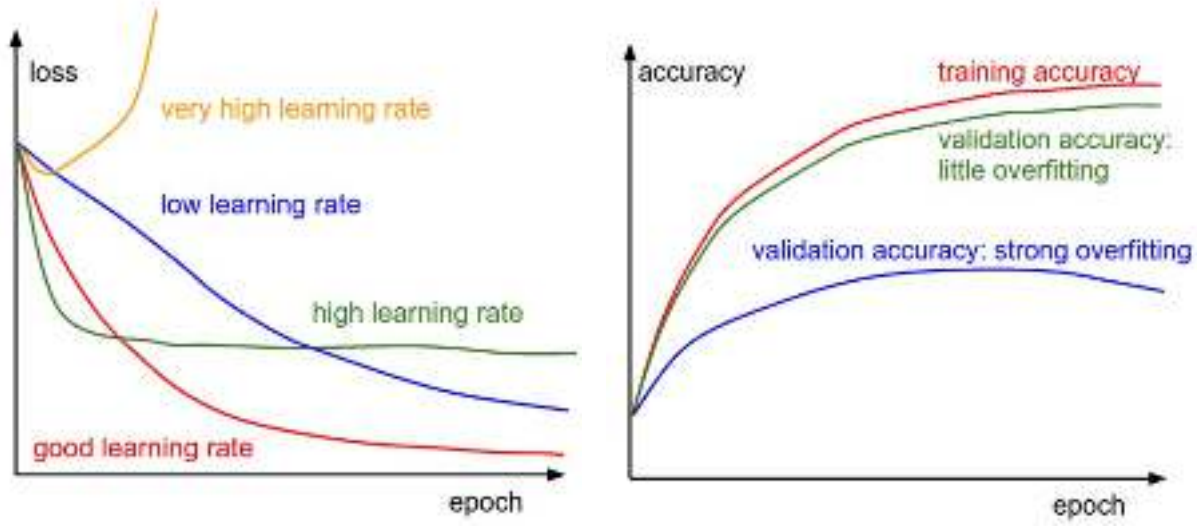
The loss function corresponds then to the average cross-entropy.

As claimed before, the machine learning models aim to construct a prediction function which minimizes the loss function. There are many algorithms which aim to minimize the loss function. Most of them are iterative and operate by decreasing the loss function following a descent direction. These methods solve the problem when the loss function is supposed to be convex. The main idea can be expressed simply as follows: starting from initial arbitrary (or randomly) chosen point in the parameter space, they allow the "descent" to the minimum of the loss function accordingly to the chosen set of directions [106]. Here we discuss some of the most known and used optimization algorithms in this field.

2.3.2 Optimization methods

The process of learning the network parameters and finding good hyperparameters have to be considered through the variation of loss value during the forward pass in training step. Hence, with a low learning rates the improvements will be linear. With high learning rates they will start to look more exponential. Higher learning rates will decay the loss faster, but they get stuck at worse values of loss. The second important quantity to track while

training a deep network is the validation/training accuracy. The plot of accuracy values during learning the model can give a valuable insights into the amount of overfitting in the learned model. The gap between the training and validation accuracy indicates the amount of overfitting [57]. Figure 2.11, plot the loss and acuracy values over different units of epochs, which measure how many times every example has been seen during training.



(a) Loss curve during training (b) training and validation accuracy curves.

Figure 2.11: Different learning rate where training and validation of a Deep CNN [57].

The Gradient Descent Algorithm

The gradient descent algorithm is the most simple and most used algorithm to find parameters for the learning model under the assumption of convexity of function to minimize. There are mainly two versions of this algorithm, the first one acts in a batch mode and the other in on-line mode. The batch mode: when we aim to minimize globally the loss function (this is why it is named batch), we first initialize randomly the parameters and we iteratively minimize the loss function by updating the parameters. This updating is done following the opposite direction of the gradient of the loss function which, locally, shows the highest slope of this function. Hence, at iteration t, the new values of the weights $w^{(t+1)}$ are estimated using the values of the weights at step t and the gradient of the loss function estimated at weight $w^{(t)}$:

$$\forall t \in \mathbb{N},\ W_{(t+1)} = W_{(t)} - \alpha \nabla \mathscr{L}\left(W_{(t)}\right), \tag{2.14}$$

where $\alpha \in \mathbb{R}_+^*$ is a positive real called learning rate. One fundamental issue is how to

choose the learning rate. If this rate is too large, than we may obtain oscillations around the minimum. If it is two small, then the convergence toward the minimum will be too slow and in same cases it may never happen.

The on-line mode: when we are dealing with large set of data, batch algorithms are not useful anymore since they are not scalable. Many works have been done to overcome this issue and to design on-line algorithms. These algorithms consider a single example at each iteration and are shown to be more efficient both in time and space complexities.

Among all the on-line algorithms, the *stochastic gradient Descent* (SGD for short) is considered as the most popular and the most used one. Many works have proved its efficiency and its scalability.

The SGD algorithm is an iterative process which acts as follows: at each iteration t, a training example (X_t, Y_t): $(x_1^t, x_2^t, \cdots, x_n^t, y^t)$ is chosen uniformly at random and is used to update the weights of the loss function following the opposite of the gradient of this function. The SGD algorithm belongs to first-order methods, i.e., those that form the parameter update on the basis of only first order gradient information. First-order methods, when used to solve convex optimization problems, have been shown to have a convergence speed, when used with large dimension problems, which can not be better than sub-linear in means of $t^{-1/2}$, [105], where t is the number of iterations. This theoretical result implies that first-order methods can not be used to solve, scalable problems in an acceptable time and with high accuracy.

Momentum is a method that helps accelerate SGD in the relevant direction. It achieves this by adding a fraction of the update vector of the past time step to the current update vector. The most popular is the method of Nesterov Momentum [96]:

$$\begin{aligned} \forall t \in \mathbb{N}, \quad Y_{(t)} &= W_{(t)} + \frac{t}{t+1}\left(W_{(t)} - W_{(t-1)}\right) \\ W_{(t+1)} &= Y_{(t)} - \alpha \nabla \mathscr{L}\left(Y_{(t)}\right), \end{aligned} \tag{2.15}$$

Nesterov momentum enjoys stronger theoretical converge guarantees for convex functions. Instead of evaluating gradient at the current position, with Nesterov momentum, the gradient is evaluated at the "looked-ahead" position.

2.4 Problem of Noise in training data

In data mining, noise has two different main sources [141]. Different types of measurement tools induce implicit errors that yield noisy labels in training data. Besides, random errors introduced by experts or batch processes when the data are gathered can produce

the noise as well. Noise of data could adversely disturb the classification accuracy of classifiers trained on this data. In the study [97], four supervised learners (naive Bayesian probabilistic classifier, the C4.5 decision tree, the IBk instance-based learner and the SMO support vector machine) were selected to compare the sensitivity with regard to different degrees of noise. A systematic evaluation and analysis of the impact of class noise and attribute noise on the system performance in machine learning was presented in [141].

The Deep CNNs use the stacking of different kinds of layers (convolution, pooling, normalization,...) that ensures the extraction of features which lead to the learning of the model. The training of deep CNN parameters is frequently done with the stochastic gradient descent 'SGD' technique [54], see section 2.3.2. For a simple supervised learning the SGD method still remains the best learning algorithm when the training set is large. With the wide propagation of convolutional neural networks, and the massive labeled data needed to train the CNNs networks, studies of the impact of noisy data was needed. A general framework to train CNNs with only a limited number of clean labels and millions of noisy labels was introduced in [132] in order to model the relationships between images, class labels and label noises with a probabilistic graphical model and further integrate it into an end-to-end deep learning system. In [110], substantial robustness to label noise of deep CNNs was proposed using a generic way to handle noisy and incomplete labeling. This is realized by augmenting the prediction objective with a notion of consistency.

Our research focused on noise produced by random errors was typically addresses a two-class classification problem: for each region in an image/video plane it is necessary to give the confidence to be salient or not for a human observer. One main contribution of this chapter is to identify how noise of data impacts performance of deep networks in the problem of visual saliency prediction. Here, to study the impact of the noise in ground truth labels, two experiments on the large data set were conducted. In the first experiment non-salient windows were randomly selected in an image plane in a standard way, just excluding already selected salient windows. Nevertheless, in video, dynamic switching of attention to distractors or to smooth pursuit of moving objects, makes such a method fail. This policy of selection of non-salient areas yields random errors. In the second experiment, cinematographic production rule of 3/3 for non-salient patches selection was used, excluding the patches already defined as salient area in all the videos frames and excluding the area where the content producers - photographers or cameramen place important scene details. The results show the increase in accuracy in the most efficient model up to 8%, all other settings being equal : the network architecture, optimization method, input data configuration.

2.5 Transfer Learning

Generally, in machine learning a simple classifier compute an output score Y from a vector X. It can be written as follow :

$$Y = f(W.X) = f(\sum_j W.X) \tag{2.16}$$

where W is a vector of weights and f is a function that converts the dot product of the two vectors into the desired output. Transfer learning techniques answer the question "How to use the vector of weights W that already trained on one problem to a different related problem?"

Transfer learning also defined as a fine-tuning techniques [6], presents a technique used in the field of machine learning that increases the accuracy of learning either by using it in different tasks, or in the same task [134] . Training CNNs from scratch is relatively hard due to the insufficient size of available training dataset in real-world classification problems. Pre-training a deep CNNs by using an initialization or a fixed feature extractor presents the heart of the transfer method. In the literature, for supervised learning with fine-tuning a variant was explored and introduced in 2006 in [45].

1. Initialize the supervised predictor (parametrized representation function $h_L(x)$ and the linear or non-linear predictor),

2. Fine-tune the supervised predicor with respect to a supervised training criterion, based on a labled training set of (x,label) pairs, and optimizing the parametres of the supervised predictor.

Gradient descent can be used for fine-tuning the weights in such "autoencoder" networks, but this works well only if the initial weights are close to a good solution. The effective way of initializing the weights is by allowing deep autoencoder networks to learn low-dimensional codes [45]. This idea work better than principal components analysis as a tool to reduce the dimensionality of data. Starting with random weights in the two networks (see figure 2.12), they can be trained together by minimizing the discrepancy between the original data and its reconstruction. Pretraining consists of learning a stack of restricted Boltzmann machines (RBMs), each having only one layer of feature detectors. The learned feature activations of one RBM are used as the "data" for training the next RBM in the stack. After the pretraining, the RBMs are "unrolled" to create a deep autoencoder, which is then fine-tuned using backpropagation of error derivatives.

In the research of Bengio et al. [134] addressing object recognition problem, the authors show that the first layers of a Deep CNN learn characteristics similar to the responses

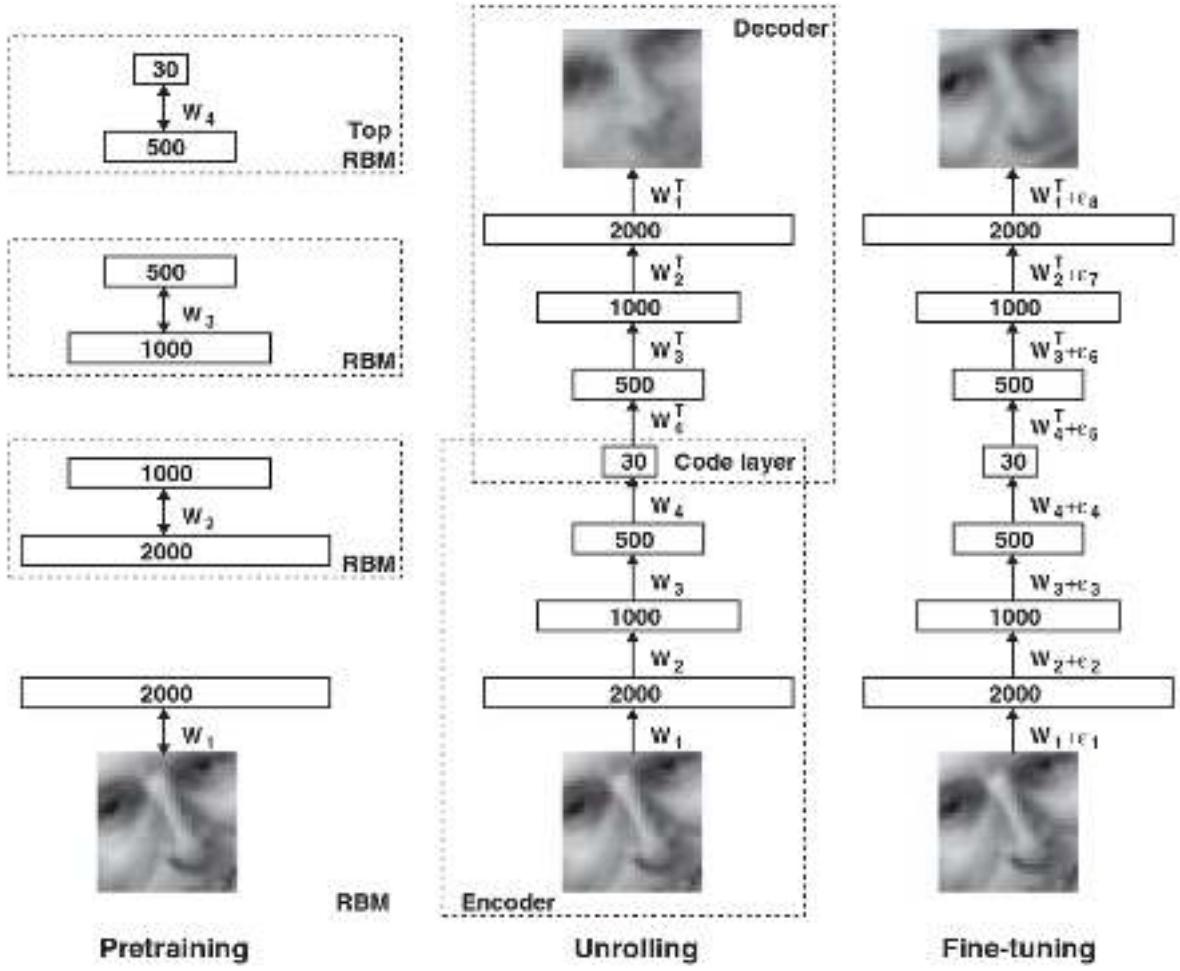


Figure 2.12: Process of transfer learning proposed by [45]. (Ref. [45])

of Gabor's filters regardless of the data set or task. Hence in their transfer learning scheme just the three first convolutional layers already trained on one training set are used for the initialization for parameter training on another training set. The coefficients on deeper layers are left free for optimization, that is initialized randomly. Several studies have proven the power of this technique [136] , [90] . Here, two famous scenarios of transfer learning with CNNs were followed :

i) using a fixed feature extractor with removing the last fully-connected layer. Here the training is fulfilled just for the linear classifier on the new dataset.

ii) Fine-tuning the weights of the pre-trained deep CNN by continuing the back-propagation [134]. Transfer learning with deep CNN shows its efficiency in different application domain such as person re-identification [33].

2.6 Saliency prediction by Deep CNNs

Saliency is useful tool that can be used in a plethora of computer vision applications such as image quality [138], superpixel [82], localization [25], retrieval [2], etc. Recently, saliency methods have been also used as the main data to predict visual fixations (scan-paths) [68]. A lot of methods to predict the saliency have been proposed in the literature. Some of them are based on low-level features by considering texture, color, intensity and orientation [52], while some others are based on high-level features or based on perceptual aspects [74]. In this section, we are interested only on methods that are based on deep learning.

Deep learning models have been used in different applications (segmentation, classification, scene understanding and so on). They have been also used to predict the saliency in image. This last decade, several saliency methods based on convolutional neural networks have been proposed in the literature. Among the first CNN model for Saliency prediction has been proposed in [69].

The basic deep learning architectures is hierarchically created with neural networks. The architectures of these networks can differ essentially by the formulation of the main problem. This formulation affects the quantity and the scheduling of convolution and pooling layers, the pooling strategies, the input data, the nature of the final classifiers and the loss functions to optimize.

Shen [115] proposes a model that approximates human gaze fixations. This model is formed by three layer sequences of "sparse coding filtering" and "max pooling", followed by a layer of linear SVM classifier to extract salient areas in images. The proposed deep learning model providing ranked "salient" or "non-salient" areas of the image, allows the learning of the relevant characteristics of the saliency of natural images, and the prediction of the eye fixations on objects with semantic content.

In Vig's work [130], the proposed learning model tackles prediction of saliency of pixels for a human visual system (HVS) and corresponds to a free-viewing visual experiment.

The learning model of the saliency of image for a specified class is defined in [117]. The challenge of this research [117] is the creation of the saliency map for each class using deep convoltional neural networks "CNN" with optimization of parameters by stochastic gradient descent. Therefore the classification problem is multi-class, and can be expressed as a "task-dependent" visual experiment, where the subjects are asked to look for an object of a given class of considered taxonomy in the images. After generating the map that maximizes the score of the specific class, the saliency map of each class is defined by the amplitude of the weight calculated from the convolution network with a single layer. In

our case, we tackle a two class classification problem: for each region in a video frame, the confidence has to be computed to belong to a "salient" class or to a "non-salient" one.

In [80], a multiresolution convolutional neural network, so called Mr-CNN, model has been proposed. The raw image is first rescaled to three different scales. Batches centered on fixation and non-fixation points with a size equal to 42x42, are extracted from the rescaled images and are used as inputs to train the proposed CNN model. Eye fixations are here used as targets.

In [75],the authors propose a multi-scale neural network architecture to predict the saliency. The input of the proposed method is the raw image decomposed into regions (segmentation). Each of these image regions has almost the same saliency value. From each considered region, three patches are then extracted (three scales), which are respectively the bounding box of the considered region, the bounding box of its neighboring regions and the whole image, and are used as input to three CNN models. In [77], an extended version has been proposed. The authors propose to boost the saliency performance by concatenating handcrafted low-level features: color and Texture (color RGB, LAB and HSV histograms, LBP histogram and the histogram of the max responses of LM filters).

In order to not predict the saliency of the whole image from patches and thus to get out the blurry filtering generally applied in this kind of method, some authors proposed fully convolutional network models, so called "end-to-end convolutional network", to predict saliency [101], [76]. This kind of models has been also applied to resolve segmentation problem [114]. In [101], the model is composed of 5 layers and predict the saliency from the image with a size of $96 \times 96 \times 3$. The authors propose to adopt the end-to-end solution as a regression problem. In [76], the authors proposed also an end-to-end model. This method uses the whole image as input and it is based on two main components: pixel-level fully convolutional and a segment-wise spatial pooling streams. The first stream aims to take into account the multi-scale properties, while the second stream aims to consider the saliency discontinuities.

In [135], a multi-scale and multi-levels model has been described. The raw image is first convoluted with some learned filters (k-means). The obtained maps are then pooled at multiple scales (four different sizes) and intermediate saliency maps are computed at multiple levels with different filters. The global saliency map is finally given summing all intermediate saliency maps and weighting it with a 2D Gaussian heat map.

In [140], the authors focus on the pooling step and proposed a global average pooling method, so-called class activation maps (CAM). This pooling step aims to produce the desired output (class) and is applied on the last convolutional layer. So, from a same

image different saliency maps can be obtained according to the object category.

In [59], the authors propose to design a CNN model for collecting eye tracking data on mobile devices. This model is composed of four inputs : left eye, right eye face images and a binary mask that provides the position of the face in the captured image.

In [70], the authors propose to combine High and low level features. The high level features are extracted from the VGG-net model, while the low level features are given by some handcrafted features based on color and Gabor filter responses. A low level distance map is then derived from the comparison of the obtained low level features and other parts of the image. The final saliency map is achieved by combining the high level features and the encoded version of the low level distance map.

A lot of works today, are devoted to saliency prediction in still images using fully convolutional network"FCN".

In [24] proposed a mixture of experts based model to predict image saliency. This model which was trained in an end-to-end manner, used global scene information in addition to local information from a convolutional neural network. The global scene information was trained on diverse categories of an eye-tracking dataset. The final saliency map is a weighted sum of the expert saliency maps.

[92] present an architectural extension to any Convolutional Neural Network (CNN) to fine-tune traditional 2D saliency prediction to Omnidirectional Images (ODIs) in an end-to-end manner. This extension present a refinement architecture that is added after the Base CNN. It takes a 3-channel feature map as input: the output saliency map of the Base CNN and the spherical coordinates per pixel as two channels.

[66] presented an approach integrating class-specific saliency maps into an end-to-end architecture to perform a weakly supervised object detection. It exploits saliency information thoroughly to boost the performance of both detection and classification. A highly confident object proposals was selected under the guidance of class-specific saliency maps. The location information, together with semantic and saliency information, of the selected proposals are then used to explicitly supervise the network by imposing two additional losses.

[71] proposed a unified deep learning framework for accurate and efficient saliency detection. The method used low-level features and high-level features which are extracted using GoogLeNe for saliency detection. The low-level features evaluate the relative importance of a local region using its differences from other regions in an image.

In [64], the authors proposed a deep CNN that predicts eye fixations and segments salient objects. The authors work on a kind of scenes with a very well distinguishable salient object and rather uninteresting background. [65] reuses an existing neural network

pretrained on the task of object recognition to predict eye fixations. [102] formulated the prediction of eye fixations as a minimization of a loss function that measures the Euclidean distance of the predicted saliency map with the provided ground truth. Despite the popularity of these models they still need a thorough study in real-life situation, which is our case.

2.7 Conclusion

In this chapter, a state-of-the-art of deep learning for visual saliency prediction was provided. We first presented the important definitions and characteristics about machine learning and especially deep convolutional networks.

Then we introduced the problem the noise in big data. Hence, a noisy data could adversely disturb the classification accuracy of learned classifiers.

We answer to the question, how transfer learning can increases the accuracy of learning and then resolve training on small data.

Finally, we provided a state-of-the-art of saliency prediction by deep CNNs. Hence, several saliency methods based on convolutional neural networks have been proposed in the literature.

In the next chapter we will present our contribution in saliency modeling using deep networks.

Part II

Deep CNNs for saliency prediction

This part describes the contribution of saliency prediction with a deep CNN. The architecture of deep CNN and the strategy of reconstruction of the saliency map are analysed here. The first chapter details the deep CNN architecture designed for the saliency prediction task. We define the classification problem for saliency prediction and propose a method to densify the response of the trained model in order to generate the final saliency map. Specific features as contrasts have demonstrated efficiency in state-of-the-art methods for saliency prediction. Second chapter resumes the use of these specific features and tests the influence of noisy data, for training a deep CNN.

Chapter 3

ChaboNet : a deep CNN designed for prediction of visual saliency in natural video

3.1 Introduction

Supervised learning techniques help with the detection of salient regions in images by predicting attractors on the basis of seen data[130]. Recent research has been directed towards the creation of a basic deep learning model that ensures the detection of salient areas. While a significant effort has been already made for building such models from still images, very few models have been built for saliency prediction in video content with supervised learning approaches [40]. Video has a supplementary dimension: the temporality expressed by apparent motion in the image plane.

The actual trend for prediction of salient areas consists in the use of supervised learning tools such as Deep CNNs. Deep CNNs were developed in Computer Vision, firstly by Yann LeCun with the LeNet [69] architecture that was used to recognize digits. Then, AlexNet[63] network has become very popular as architecture for visual recognition tasks. It has a very similar architecture to LeNet, but is larger in terms of number of convolutional filters, deeper, and featured Convolutional Layers are stacked on top of each other. In prediction of visual saliency, the deep CNNs are becoming popular as well [130], [122], [115], [117] .

Deep learning architectures, which have recently been proposed for the prediction of salient areas in images, differ essentially by the quantity of convolution and pooling layers, the input data, pooling strategies, the nature of the final classifiers, the loss functions to optimize and the formulation of the problem.

3.2 General approach

In the variety of predictors of visual attention in images and video we are interested in predicting "static" visual attention, which means that for each pixel (x, y) in image plane depicting a visual scene, we aim to predict its importance or salincy $S_M(x, y)$. Nevertheless, unlike classical methods for prediction of static saliency maps, in our supervised learning framework we propose a two step approach:

*Step*1: here we wish to roughly delimit, "spotify" regions-of-interest in the image plane. Hence the problem consists in the prediction of saliency not of a single pixel but of a whole region. Without any pre-segmentaiton of the image plane we work with regular grid of squared patches P_i.

*Step*2: Then on the basis of densly sampled patches we can interpolate the saliency map $S_M(x, y)$ for each pixel(x, y).

The overall block diagram of proposed approach for saliency prediction is depicted in figure 3.1 . After a various number of training and validation iterations, a trained deep CNN model was obtained. This trained model ensures the prediction of saliency probability for each regions that are obtained by dense sampling the input image frame. Using the responses of trained model on sampled patches, we interpolate the final saliency map.

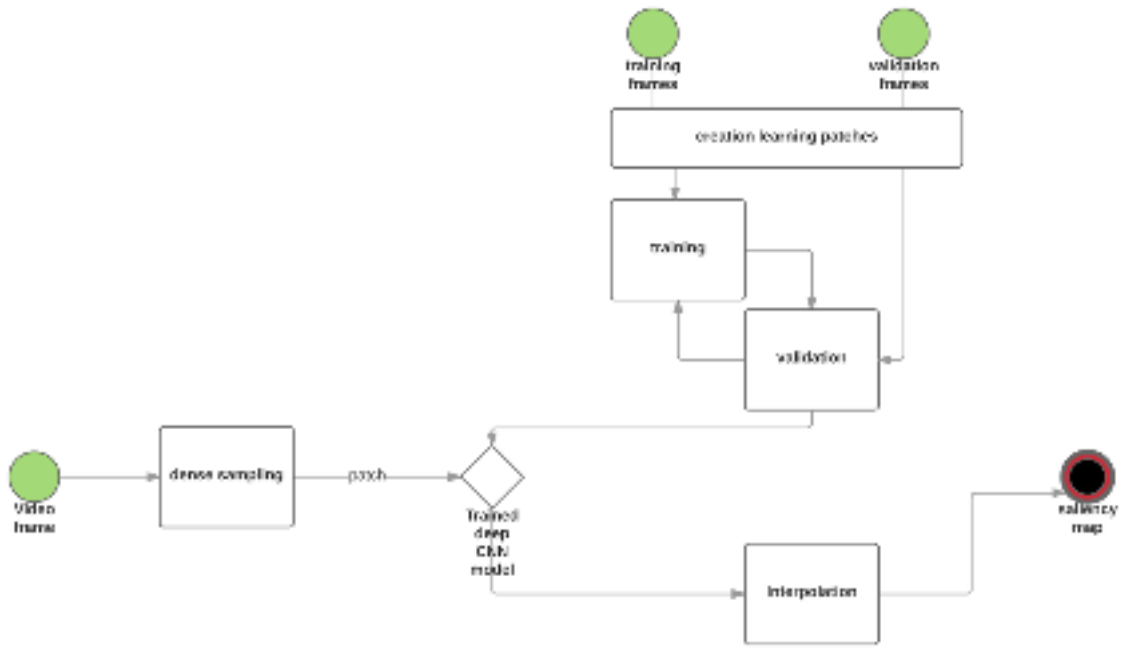


Figure 3.1: Overall block diagram of proposed approach for saliency prediction.

3.3 Policy of data set creation: salient and Non-salient patches

Whatever is the architecture of a Deep CNN for saliency prediction, selection of a traning dataset which would contain as less noise as possible is the must. The training set has to be built to comprise salient and Non-salient regions in video frames. The ground-truth for saliency here are the Gaze Fixation Density Maps (GFDM). They are built upon gaze fixations of a cohort of subjects recorded during a psycho-visual experiment. We formalize it in subsection 3.3.1.

For salient patches extraction the intuition is clear: we need to extract patches in the video frames where the GFDM has strong values. For Non-salient patches extraction,the situation is more complex. Due to the distractors and visual fatigue, the areas in a given video frame which are salient can become Non-salient in the next frame. Thus the noise is introduced in the training set of Non-salient patches. We thus proposed a strategy based on video production rules which will allow to avoid the noise as much as possible. It is presented in subsection 3.3.2.

3.3.1 Salient patches extraction

In the following equations bold variables will denote vectors. A squared patch $\mathbf{P}$ of size $s \times s \times k$ ($s = 100$ adapted to the spatial resolution of standard definition (SD) video) in a video frame is defined as a vector in $R^{s \times s \times k}$. Here k stands for the quantity of primary feature maps serving as an input to the deep CNN. In case when conventional RGB planes are used as input data for the network, then $k = 3$; if supplementary data layer, such as motion is added, then $k = 4$.

Patch saliency is defined on the basis of its interest for subjects. The interest is measured by the magnitude of a GFDM built upon gaze fixations for each video frame. GFDMs are built by the method of Wooding [131]. Such a map $S_g(x, y)$ represents a multi-Gaussian surface. Each Gaussian is centered on a gaze fixation point. Then the Gaussians are summed up and the surface is normalized by its global maximum.

A binary label is associated with each patch $\mathbf{P}_i$ using equation (3.1).

$$L(\mathbf{P}_i) = \begin{cases} 1 & if \quad S_g(x_{0,i}, y_{0,i}) \geq \tau_J \\ 0 & otherwise \end{cases} \tag{3.1}$$

with $(x_{0,i}, y_{0,i})$ the coordinates of the patch center in the image plane. A set of thresholds is selected starting by the global maximum value of the normalized GFDM and then

relaxing threshold values as in equation (3.2):

$$\begin{cases} \tau_0 = \max(S_g(x,y))(x,y) \in D \\ \tau_{(j+1)} = \tau_j - \delta\tau_j \end{cases} \tag{3.2}$$

Here D is the image definition domain, $0 < \delta < 1$ is a relaxation parameter, $j = 0, \cdots, J$, and J limits the relaxation of saliency. It was chosen experimentally as $J = 5$, while $\delta = 0.04$. In complex scenes several details or objects can attract human attention. Thus the map $S_g(x,y)$ can contain several local maxima. In order to highlight them, morphological erosion with 3x3 structuring element was applied to $S_g(x,y)$.

Figure 3.2 summarizes different steps to select salient patches. Firstly, the GFDMs were computed, then the operation of erosion was applied. The illustration is given at a frame from HOLLYWOOD [1] dataset. Patches centered on local maxima with saliency values satisfying the equations (3.1), (3.2) are selected as salient. Retained salient patches should be distanced at least by $(\frac{1}{2} \times s)$. Non-salient patches extraction is described in section 3.3.2.

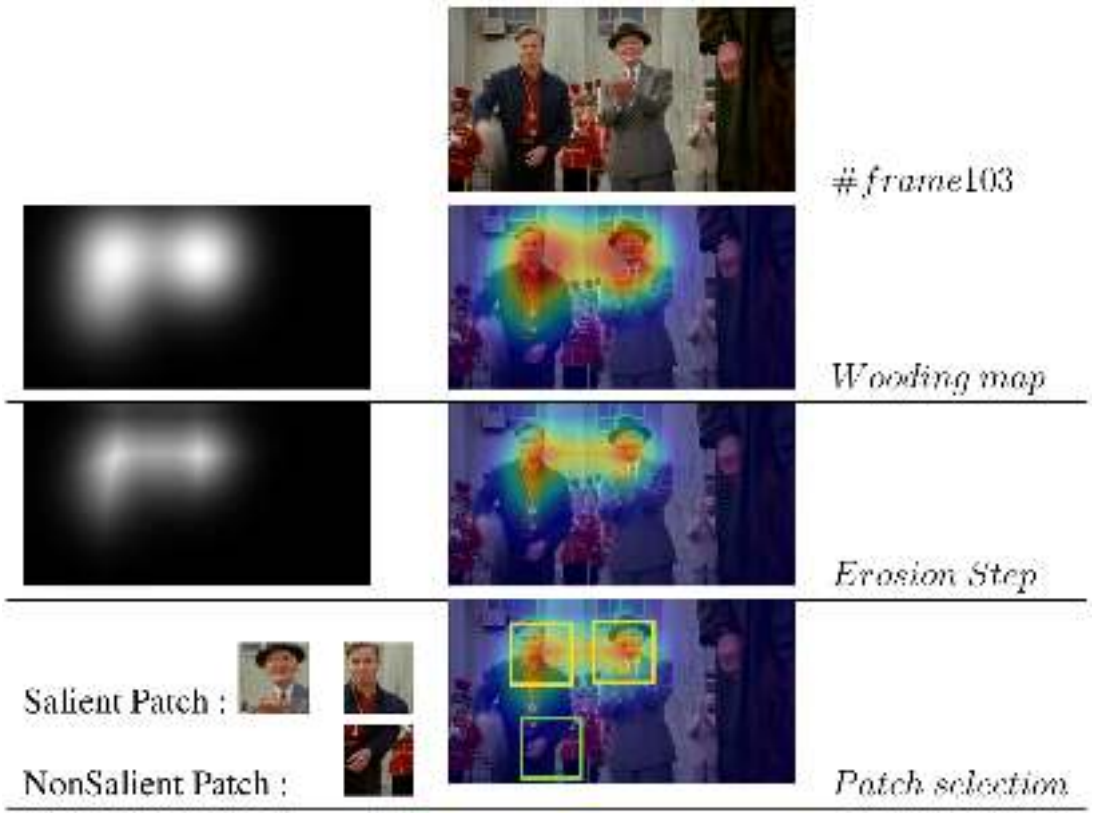


Figure 3.2: Policy of patch selection : example and steps (HOLLYWOOD[88] [89] data set 'actioncliptest00003'.

[1] available at http://www.di.ens.fr/∼laptev/actions/hollywood2/

3.3.2 Non-salient patches extraction

A Non-salient patch is a squared region in the image plane which is not supposed to attract human gaze. In the following we will expose two methods of selection of Non-salient patches for training of "Non-salient" class in our supervised-learning framework for saliency prediction.

Method 1

Let us, for a given video frame at time t denote $SP(t)$ a set of pixels belonging to selected salient patches and $SP(t) \cup \overline{SP}(t) = D(t)$. Then any patch $\mathbf{P}_j$ with all its pixels in $\overline{SP}(t)$ can be considered as a Non-salient. Therefore, the *first method* of selection of Non-salient patches consists in random selection of patch centers in $\overline{SP}(t)$ that verify the two following conditions. For each selected patch $\mathbf{P}_j$:

i) pixels of the selected patch are not in $SP(t)$;

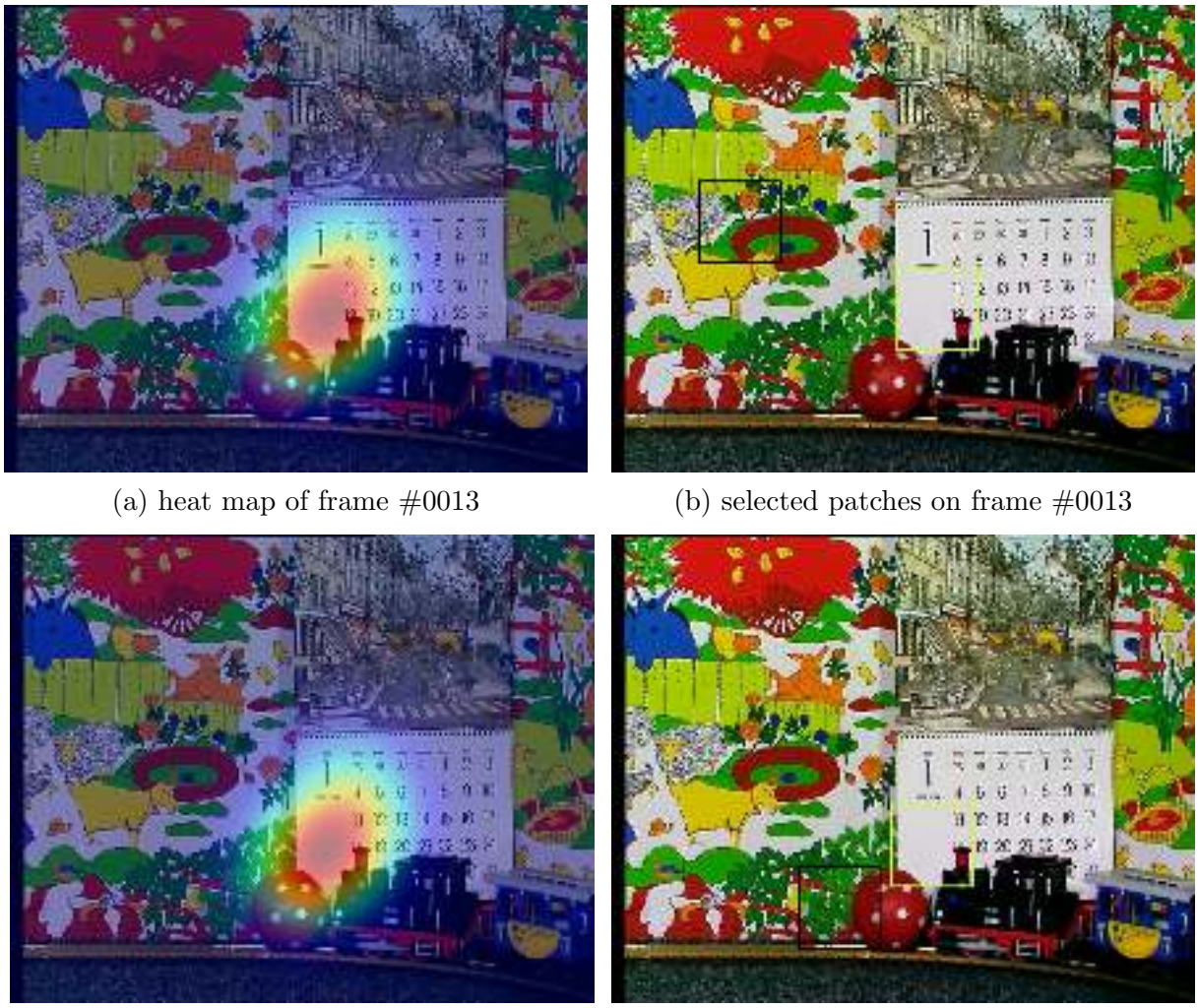
ii) The intersection for any two selected Non-salient patches $\mathbf{P}_j$ and $\mathbf{P}_k$, is empty.

The first condition guaranties that Non-salient patches correspond to the area of the current video frame, where the GFDM $S_g(x, y)$ values are low relatively to the condition 3.1 The second condition ensures a large spread of selected Non-salient patches in the image plane.

An illustration of selected salient and Non-salient patches in a video frame generated by the first method is presented in figure 3.3. "salient" patches are presented by green square and "Non-salient"by black one.

In the bottom-up saliency definition, local contrasts can invoke human gaze. Here, when analysing selected patches, we can state that Non-salient patches can contain parts of contrasted objects (a "Non-saillent" patch in figure 3.3 (b) is selected on a contrasted background. In figure 3.3 (c) it is selected even on the moving object (red ball).). For saliency prediction tasks, the main difference when designingof supervised learning approches vs bottom-up methodes is that Non-salient patches can contains a contrasted area. The former exploit the interest of subjects in the visual content expressed by gaze fixation density maps only, while the latter are purely stimuli(/image)- driven.

Nevertheless, such a straight-forward method for Non-salient patches extraction yields a noise in the training data. According to our observations, in video areas of high saliency can change in-between frames, this is due to the distractors. We namely have observed such a phenomenon in the intentioally degraded content that we produced for assessment of patients with neuro-degenerative diseases. The focus of attention of healthy subjects change when they observe the degraded sequence and especially during an appear of unusual intentionnally degraded area in video frame (see chapter 6). This is illustrated in figure 3.4 below:



(a) heat map of frame #0013

(b) selected patches on frame #0013

(c) heat map of frame #0014

(d) selected patches on frame #0014

Figure 3.3: Extraction of Non-salient patches by random selection in the Non-salient area of a video frame: Random selection of Non-salient patches on successive frames of SRC07 video IRCCyN [16].

(a) frame #388

(b) frame #399

(c) frame #533

Figure 3.4: Change of focus of attention due to distractors : Switched saillent object (degraded elephent and car) on degraded sequence create noise (heat map on frames #388, #399 and #533).

Such a change yields the erros in selection of Non-salient patches, indeed, as the focus of attention is shifted in frame at $t+1$ a non salient patch can be selected on the object which was salient in the frame at t. Thus selected a Non-salient patch $\mathbf{P}_j$ is then a noise data in a "Non-salient" class of the training set.

The problem of noise in training data and its influence on class-prediction accuracy in Deep learning is one of the open and urgent problems of the machine learning that community is facing now [55], [60] , [81]. In the context of saliency prediction in visual content, it is important not only for video, but also for the approaches on co-saliency detection from collections of images [137]:

To our best knwoledge this problem has not been adressed yet in the context of prediction of visual saliency. Hence to overcome this particular noise generation, we propose a second method for the extraction of Non-salient patches based on visual content production rules.

Method 2

According to the rule of thirds in produced and post-produced digital visual content, the most interesting details of the image or of a video frame have to cover the frame center and the intersections of the three horizontal and vertical lines that divide the image into nine equal parts [84].

Let $(x_{0,i}, y_{0,i})$ be the coordinates of the center of the patch $\mathbf{P}_i$, $width$ is the width size of the video frame and $height$ is its height size. Let us denote by SP the set of all pixels belonging to salient patches selected as described in section 3.3.1. To exclude such pixels and the area-of-interest, the one-fifth band of the frame was chosen starting from its border. Then Non-salient patch centers are randomly selected in this area. Hence the generated coordinates $(x_{0,i}, y_{0,i})$ of i-th Non-salient patch satisfy the following conditions:

$$
\begin{aligned}
&[(x_{0,i}, y_{0,i}) \notin SP] \wedge \\
&\begin{cases}
[0 \leq x_{0,i} < \frac{width}{5}] \wedge [0 \leq y_{0,i} < height] \\
or[((width - \frac{width}{5}) \leq x_{0,i} < width) \wedge (0 \leq y_{0,i} < height)] \\
or[((\frac{width}{5}) \leq x_{0,i} < (width - \frac{width}{5})) \wedge (0 \leq y_{0,i} < \frac{height}{5}))] \\
or[((\frac{width}{5}) \leq x_{0,i} < (width - \frac{width}{5})) \\
\qquad \wedge (height - \frac{height}{5}) \leq y_{0,i} < height)]
\end{cases}
\end{aligned}
\tag{3.3}
$$

Schematically, the center of Non-salient patch should be in the bluefish area shown in figure 3.5. The yellow lines depict the lines of interest.

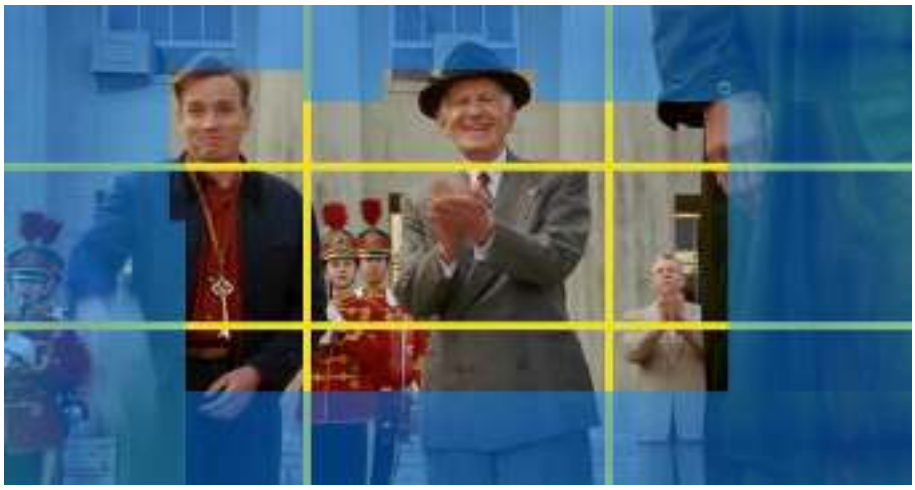
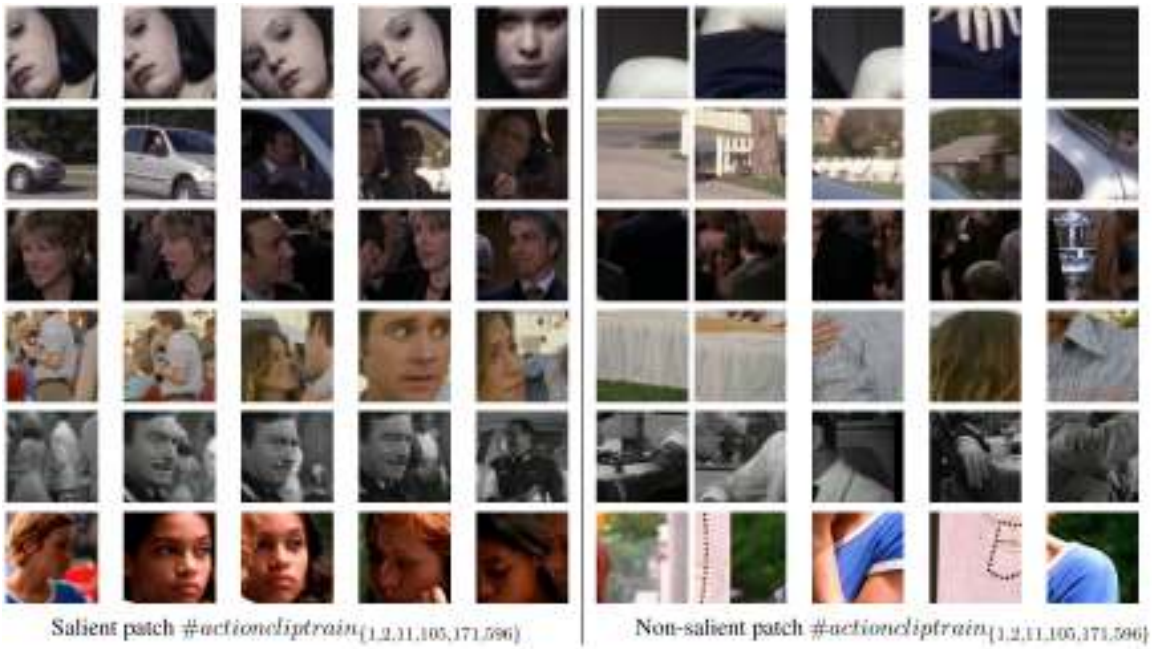


Figure 3.5: Space of selection of Non-salient patches 'actioncliptest00003'.

The table 3.1 below presents the group of salient patches on the left and Non-salient patches on the right. The rows contain some examples taken from frames of a set of video sequences "actioncliptrain" from the HOLLYWOOD[2] data set. Once more we note that Non-salient patches can contain rather strong contrasts (e.g. as in the first row on the right), but these patches have not attracted visual attention of subjects and are not situated in the area-of-interest acccordingly to the thirds rule.

Table 3.1: Training data from HOLLYWOOD data set



[2]available at http://www.di.ens.fr/∼laptev/actions/hollywood2/

3.4 Deep Convolutional Neural Network for visual saliency: ChaboNet

In this section, the proposed architecture ChaboNet for the visual saliency prediction problem is presented. As the purpose is in predicting visual saliency in video, specific features which are added to conventional RGB pixel values are described first. Then the architecture in terms of layers is presented. The implementation of ChaboNet is realized on the basis of Caffe framework [54].

3.4.1 A specific input data layer

When adressing visual attention prediction in video, the sensitivity of HVS to motion has to be taken into account [10]. Indeed in classical bottom-up saliency prediction models, the sensitivity of HVS to motion in a dynamic scene is modeled by residual motion [87]. Human observers accommodate to the global motion in a visual scene, such as camera motion, and are attracted by specific local motions of objects. They first execute a saccade to a moving target and then continue with the "smooth pursuit" or visual tracking [107] keeping focus-of-attention on it. Local motion, i.e. motion of the target is expressed by residual motion relatively to the camera motion observed in the image plane [87]. The global motion in the plane of video frames expresses camera motion. To compute residual motion, the approach described in detail in next chapter 4.2.1 was followed. Here a pixel-wise motion field is computed by an optical flow method first. Using the dense motion field vectors as raw measures, the affine linear model of global motion is estimated by RANSAC algorithm [28]. Finally, the residual motion is the vector - difference between the initial motion vector and the one generated by the estimated affine model. As motion features, the squared $L2$ norm of residual motion vectors in each pixel in a video frame, normalized by its maximum in the frame, is used.

The composition of the input layer of the CNN is illustrated in figure 3.6. Here for each patch the input layer is composed of three color channel values and the residual motion feature map. Due to this configuration, the model is called "ChaboNet4k" in contrast to "ChaboNet3k", where only color channel values are used.

3.4.2 The ChaboNet network architecture design

ChaboNet architecture was designed for the two-class classification problem: prediction of category of a patch in a given video frame as salient or Non-salient. We aimed i) to preserve a reasonable deepness and ii) to remain comparable in the number of layers with

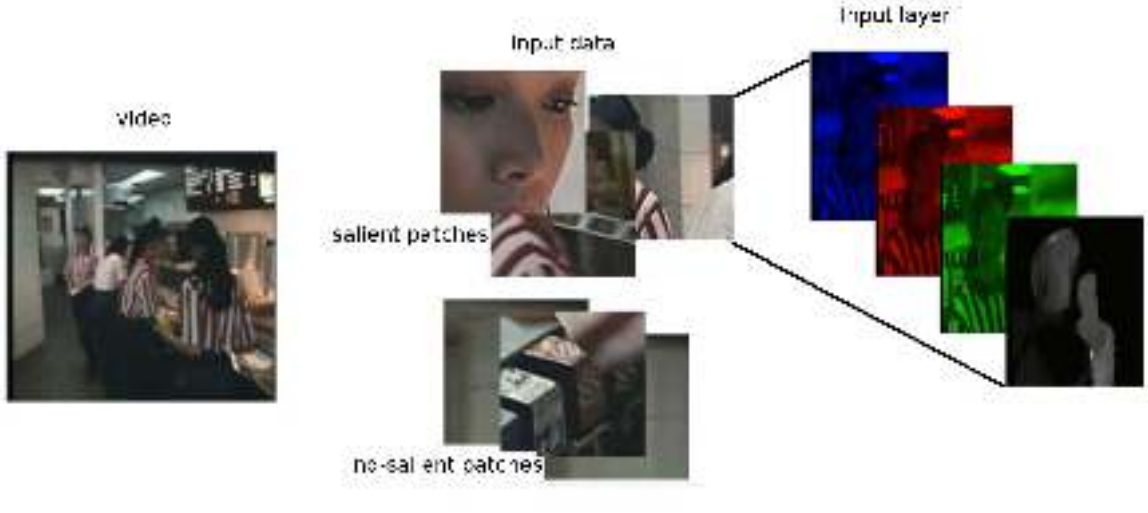


Figure 3.6: Input data layer : different features to ingest in the network.

a quite efficient network Alexnet [63]. The ChaboNet architecture is summarized in figure 3.7.

As in the majority of Deep CNN architectures designed for image classification tasks [54], ChaboNet is composed of a hierarchy of patterns. Each pattern consists of a cascade of operations, followed by a normalization operation in some cases. The cascading of linear and nonlinear operations successively produces high-level features. They are trasmitted via a fully connected layer to the deepest layer which is a soft-max classifier. It assigns the confidence for each patch to be salient or not. Due to quite a limited size of input patches three patterns were proposed in this architecture. The pattern P^1 below is a usual combination of convolution, pooling and non-linear layers, P^2 and P^3 have the same structure. The whole network can be detailed as follows.

$$\begin{aligned} &Pattern\ P^1 : \\ &\qquad Input \xrightarrow{convolution} Conv^1 \xrightarrow{pooling} Pool^1 \xrightarrow{RELU} R^1 \\ &Pattern\ P^p :\ with \quad p \in \{2,3\} \\ &\qquad N^{p-1} \xrightarrow{convolution} Conv^p \xrightarrow{RELU} R^p \xrightarrow{convolution} Conv^{pp} \xrightarrow{RELU} R^{pp} \xrightarrow{pooling} Pool^p \end{aligned}$$

The normalization operation was added after the patterns P^1 and P^2 only, as after the pattern P^3 the features are quite sparse. The architecture of ChaboNet is depicted in figure 3.7. The features after convolution layers are presented for the example image from figure 3.6. It can be seen that the first layer of the network performs more as low-pass filters and deeper the convolution layer is more "high-pass" effect is observable.

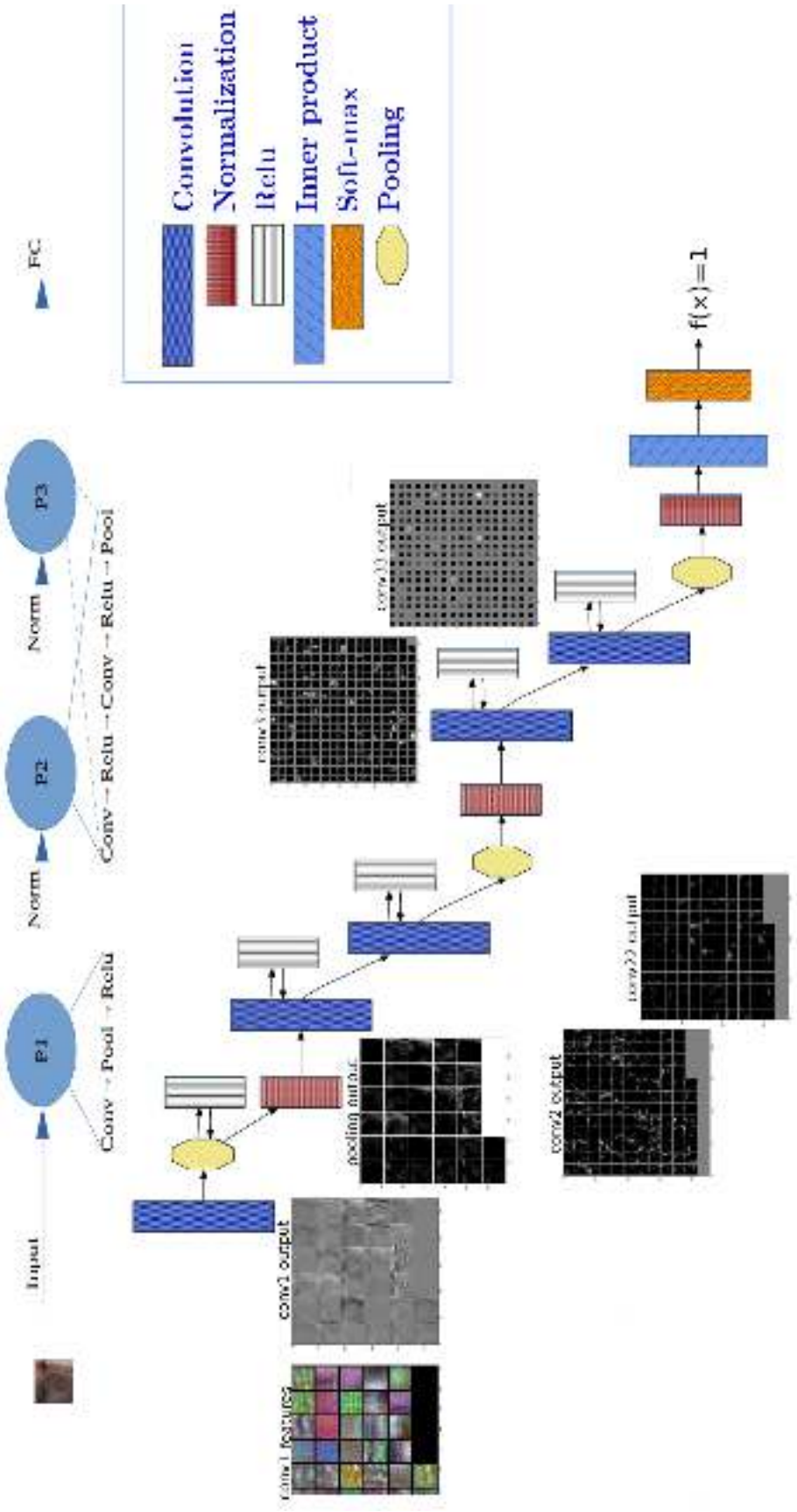


Figure 3.7: Architecture of video saliency convolution network 'ChaboNet'.

Inspired by literature as [63], [115] where the size of convolution kernels is either maintained constant or is decreasing with the depth of layers, in ChaboNet network, 32 kernels were used with the size of 12×12 for the convolution layer of the first pattern P^1. In the second pattern P^2, 128 kernels for each convolutional layer were used. In P^2 the size of the kernels for the first convolutional layer was chosen as 6×6 and for the second convolution layer, a kernel of 3×3 was used. Finally, 288 kernels with the size of 3×3 were used for each convolution layer of the last pattern P^3. This allows a progressive reduction of highly dimensional data before conveying them to the fully connected layers. The number of convolution filters is growing, on the contrary, to explore the richness of the original data and to highlight structural patterns. For the filter size, several tests were made with the same values as in AlexNet [63], Shen's network [115], LeNet [69], Cifar [61] and finally, the size of 12×12 was retained in the first layer of the pattern P^1 as it yielded the best accuracy in saliency prediction problem.

Figure 3.8 summarizes the parameters used for each layer of the three patterns.

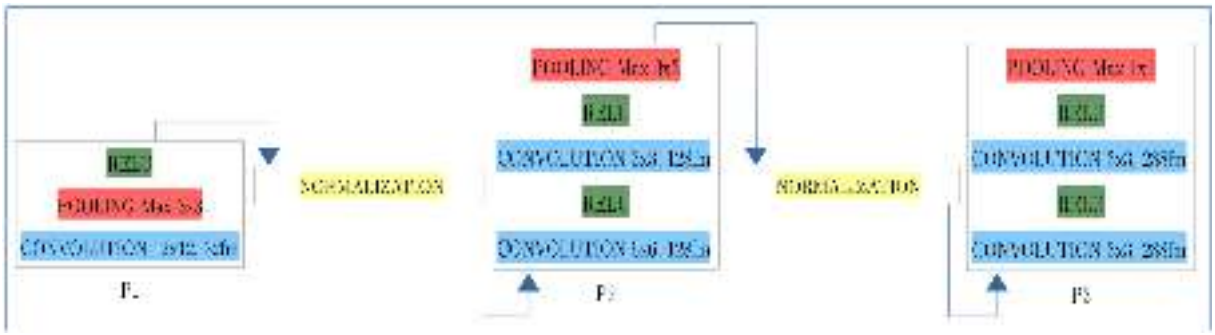


Figure 3.8: Detailed setting of each layer of 'ChaboNet' network.

3.4.3 Visualization of features

It is interesting to visualize the purely spatial features computed by the designed CNN in case when the network is configured to predict saliency only with primary RGB values. As the feature integration theory states, the HVS is sensitive to orientations and contrasts. This is what we observe in features going through layers of the network. The output of convolution layers (see figures 3.9, 3.10 and 3.11) yields more and more contrasted and structured patterns. In these figures *convi* and *convii* stand for consecutive convolution layers without pooling layers in between.

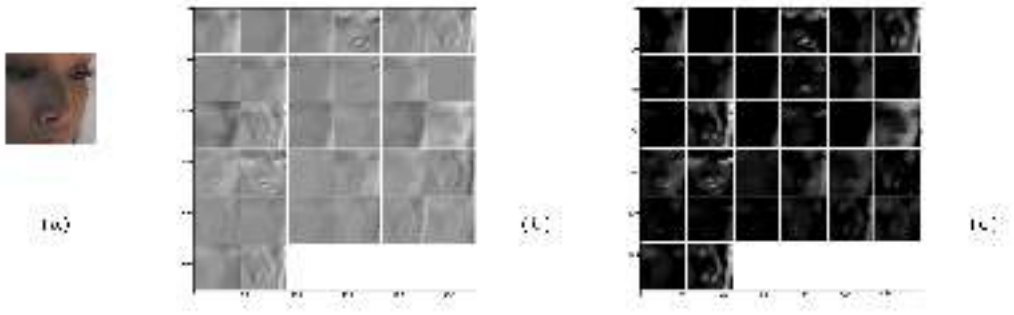


Figure 3.9: (a) Input patch, (b) the output of first convolution layer and (c) the output of the first pooling layer.

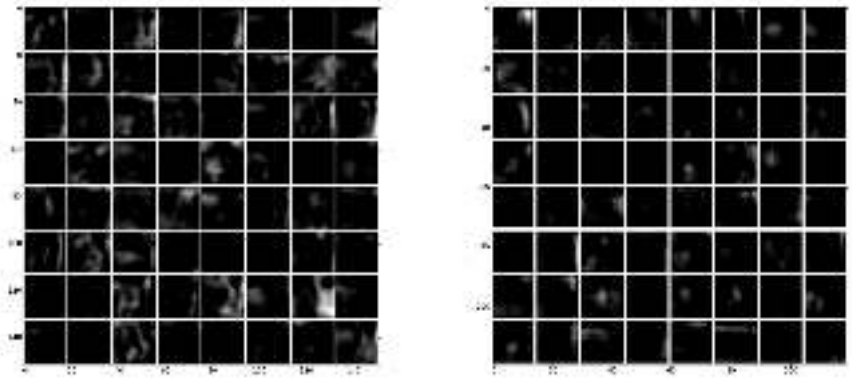


Figure 3.10: The output of the 2nd convolution layer, ' Conv2' and 'Conv22'.

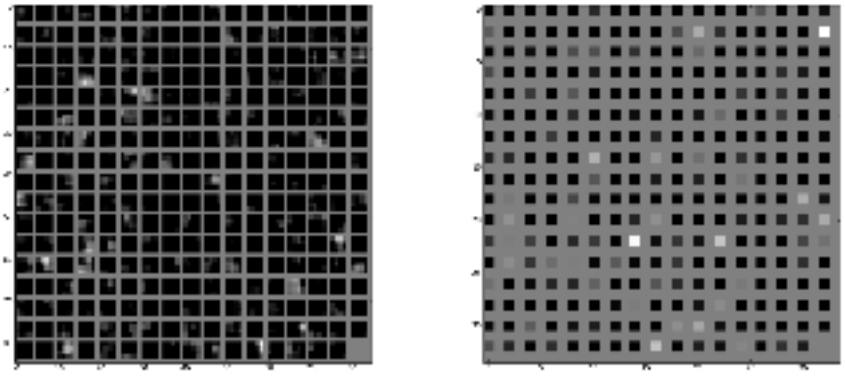


Figure 3.11: The output of the third convolution layer, ' Conv3' and 'Conv33'.

3.4.4 Training and validation of the model

To solve the learning problem and to validate the network with the purpose to generate a robust model for salient area prediction, the solver of Caffe [54] is repetitively optimizing the network parameters in a forward-backward loop. The optimization method used is the stochastic gradient descent 'SGD' with a simple momentum. Indeed, in [120] the authors explain the necessity of momentum method, which allows for avoiding of oscillations of a simple gradient descent method in Deep networks optimization. In is caracterized by an introduction of a specific term - velocity. It is a technique to accelerate gradient descent by accumulating a velocity vector in directions of persistent reduction in the objective function across iterations. The method is expressed by the following equation.

$$\begin{cases} V_{i+1} = m_o \cdot V_i - \gamma \cdot \alpha \cdot W_i - \alpha \cdot \langle \frac{\partial L}{\partial W} | W_i \rangle_{D_i} \\ W_{i+1} = W_i + V_{i+1} \end{cases} \tag{3.4}$$

With W convolution coefficients, V is the velocity vector, $\alpha = 0.001$- is a fixed learning rate, $m_o = 0.9$ - is a momentum coefficient, $\gamma = 0.00004$ is the weight decay. The initial value of the velocity V_0 was set to zero. These parameter values are inspired by the values used in [54] with the same fixed learning rate and show the best performances on a large training dataset. Further in the manuscript we will come back to the algorithm and study different ways of its initialization. In the present chapter the initialization of convolution coefficients is realized randomly according to Gaussian distribution as proposed in [54].

The parameterization of the solver requires also setting the number of iterations at training step. The number of iterations was defined accordingly to the equation (3.5):

$$iterations\_numbers = epochs \times \frac{Total\_images\_number}{batch\_size} \tag{3.5}$$

here *batch_size* represents the number of images for each network switching, *epochs* is the number of times the totality of the dataset is switched by the network. We will study this parameter in the experimental part of the present chapter.

3.5 Generation of saliency map

The saliency map of each frame I of the video is constructed using the output value, for each patch, of the trained deep CNN model. We proposed to interpolate sparse classification results. The soft-max classifier that takes the output of the inner product

layer as input, gives the probability for a patch to belonging to the salient class. Function defined in equation 3.6 presents a generalization of the logistic function that compresses a vector $\mathbf{U}$ of arbitrary real values of dimension d to a vector of the same dimension but with actual coordinate values in the range $(0, 1)$.

$$\phi(\mathbf{U})_q = \frac{e^{u_q}}{\sum_r e^{u_r}}, r = 1, ..., d \tag{3.6}$$

Hence, from each frame I local regions having the same size as training patches (here $s = 100$) are selected in a raster-scan scanning process. The output value of the soft-max classifier with regard to the salient class on each local region defines its degree of saliency. If the score is assigned to the center of each patch, a sparse saliency map is obtained $M(x, y)$. It has a non-zero values only in the center of patch (x_0, y_0). In a scanning process densely sampled, with a stride of 5 pixels, local regions were classified. Then score values assigned to the centers were interpolated with Gaussian filters: in the center of each local region, a Gaussian $G(x, y)$ was applied with a pick value of $\frac{A \times M(x_0, y_0)}{2\pi\sigma^2}$. The A-parameter value was experimentally choosen as 10. The spread parameter σ was fixed as a half-size of the patch. For each pixel in the image plane the Gaussians were summed-up. Finally the map was normalized by saliency peak as in Wooding method for GFDM (see section 1.3.1).

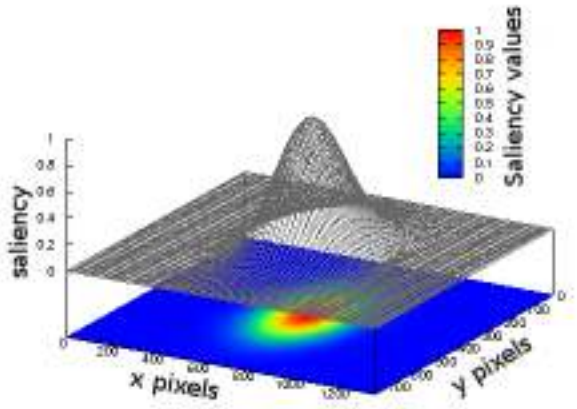


Figure 3.12: Psycho-visual 2D Gaussian depending to the fovea area on the local region center predicted as salient.

Algorithm 2 Predict saliency map (frame_RGB, frame_RMotion)

Require: frame_RGB : RGB frame I of the video,
frame_RMotion : residual motion map corresponds to I frame
Ensure: saliency map tab_saliency
begin
$frame\_4k = concatenate(frame\_RGB, frame\_RMotion, 2)$
$SizePATCH = 100$
$STRIDE = 5$
$kernelXY = Matrix\_Zero[Height(frame\_4k)][Width(frame\_4k)]$
numCores = number of CPU processor
for each process Pi in numCores **do**
$arrayRes = (delayed(getPatchClassification)(patch)$
$for(x, y, patch) in slidingWindow(frame\_4k, STRIDE, (SizePATCH, SizePATCH)))$
end for
positions = Research (positions in arrayRes> 0.0)
for x, y in positions **do**
for probI in range $(0, int(10 \times arrayRes[x][y]))$: **do**
$kernelX = getGaussianKernelWithCenter(int(HEIGHT), x \times STRIDE + int(SizePATCH/2), sigma)$
$kernelY = getGaussianKernelWithCenter(int(WIDTH), y \times STRIDE + int(SizePATCH/2), sigma)$
$kernelXY = kernelXY + kernelX \times kernelY.transpose()$
end for
end for
$kernelXY = kernelXY \times (1/maxVal(kernelXY)) \times 255$
save saliency map : $tab\_saliency = saveImage(kernelXY)$

Algorithm 3 procedure getPatchClassification (Patch)

Require: *Patch size* 100×100
Ensure: $Salient - prob =$ *probability of Patch saliency*
transformer = caffe.io.Transformer()
transformer.set-transpose()
transformer.set-mean()
transformer.set-raw-scale()
transformer.set-channel-swap()
net.blobs['data'].data[...] = transformer.preprocess('data', Patch)
out = net.forward()
return Salient-prob = out[prob(Patch)] * Label[out[prob(Patch)]];

Algorithm 4 procedure getGaussianKernelWithCenter(length, center, sigma):

Require: length, center, sigma
Ensure: Compute the gaussian kernel of saliency : gaussianKernel
$auxKernel = cv2.getGaussianKernel(length * 3, sigma, cv2.CV\_32F)$
$gaussianKernel = auxKernel[length + (length/2 - center) : 2 \times length + (length/2 - center), 0 :]$

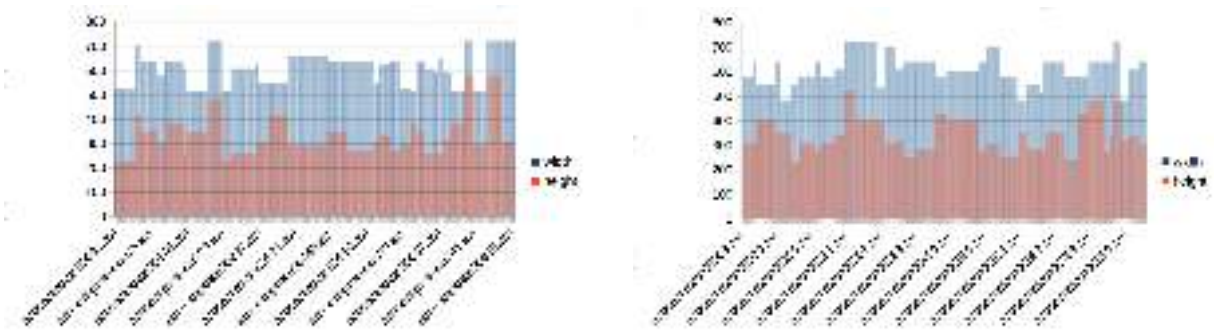
return gaussianKernel

3.6 Experiments and results

3.6.1 Data sets

To learn the model, HOLLYWOOD[88] [89] data set with approximately 20 hours of recordings in total was used.

The HOLLYWOOD database contains 823 training videos and 884 videos for the validation step. Video resolution are from 480×320 to 720×576 at $24 - 25 fps$. The distribution of spatial resolutions of videos are shown in figure 3.13. The number of subjects with recorded gaze fixations varies according to each video with up to 19 subjects. The spatial resolution of videos varies as well. Despite the discrepancy of these parameters, we use it for model building as it is the only large-scale video database with recorded gaze fixations. The HOLLYWOOD dataset contains 229825 frames for training and 257733 frames for validation. From the frames of the training set, 222863 salient patches and 221868 Non-salient patches were extracted. During the validation phase, 251294 salient patches and 250169 Non-salient patches were used respectively. The distribution of the data between "salient" and "Non-salient" classes is presented in table 3.2.



(a) training step (b) validation step

Figure 3.13: Histogram of video resolutions ($W \times H$) of "HOLLYWOOD" database in training and validation step.

Table 3.2: Distribution of learning data: total number of salient and Non-salient patches selected from each database.

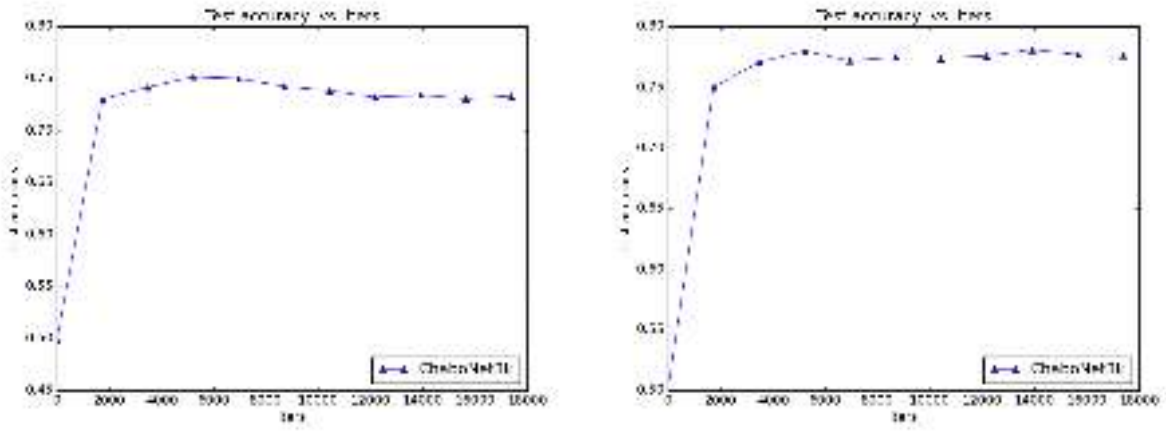
data sets		training step	validation step
HOLLYWOOD	SalientPatch	222863	251294
	Non-salientPatch	221868	250169
	total	444731	501463

3.6.2 Evaluation of patches' saliency prediction with deep CNN

The network was implemented using a powerful graphic card Tesla K40m and processor (2×14 cores). Therefore a sufficiently large amount of patches, 256, was used per iteration (see the *batch_size* parameter in equation (3.5)). After a fixed number of training iterations, a model validation step was implemented: here the accuracy of the model at the current iteration was computed on the validation data set, we call it "Test accuracy" as mentionned in the figure 3.15. In the following we first evaluate our proposal of patch selection for training with filtering of noise in training data (*Method* 2) against random patch selection (*Method* 1), see section 3.3.2.

Evaluation of Noise filtering in the training set for Non-salient patches

In figure 3.14 below, the curves of accuracy were shown on validation dataset a) for selection of training Non-salient patches by random sampling with *Method* 1, and b) when Non-salient patches are selected according to our proposed *Method* 2 using the thirds rule.



(a) Non-salient patches by random sampling (b) Non-salient patches using the rule of third

Figure 3.14: Influence of Non-salient patches selection method on resulting accuracy. a)Random selection of patches; b) Selection of patches accordingly to 3/3 rule.

It is clear, that the filtering of noisy data in training datatset in our problem of prediction of saliency of patches in video frames, allows to increase classification accuracy. We also summarize these results in terms of peak an mean statistics in the table 3.3 below. The proposed method of filtering noise in traning data in Non-salient class yields the increase of max and mean accuracy of more than 7%.

Table 3.3: The accuracy results with two methods of Non-salient patch extraction: a) Random Sampling in Non-salient area; b) Selection accordingly to 3/3 rule

	ChaboNet3k with random sampling	ChaboNet4k with 3/3 rule selection
$min_{(\#iter)}$	$49.8\%_{(\#0)}$	$50.11\%_{(\#0)}$
$max_{(\#iter)}$	$75.1\%_{(\#5214)}$	$77.98\%_{(\#5214)}$
$avg \pm std$	$71.6\% \pm 0.072$	$77.30\% \pm 0.864$

In the following experiments we thus retain the second method for selection of Non-salient patches: the 3/3 rule.

Evaluation of motion features

To evaluate our deep network and to prove the importance of the addition of the residual motion map, two models were created with the same parameter settings and architecture of the network: the first one contains R, G and B primary pixel values in patches, denoted as $ChaboNet3k$. The $ChaboNet4k$ is the model which uses RGB values and the normalized energy of residual motion as input data. Figure 3.15 illustrates the variations of the accuracy along iterations of all the models tested for the database "HOLLYWOOD". Peak and mean accuracy values are presented in table 3.4).

The results of learning experiments on HOLLYWOOD data set yield the following conclusions:

i)When adding residual motion as an input feature to RGB plane values, the accuracy is improved by almost 2%.

ii) The accuracy curve (figure 3.15 (a)) and the corresponding loss curve (figure 3.15 (b)) show that the best trained model reached 80% of accuracy with the smallest loss (at the iteration #8690 see table 3.4). Thus, it does not present an over-fitting situation.

The model obtained after 8690 iterations is used to predict saliency on the validation set of this database, and to initialize the parameters when learning with transfer on other used data sets in the Chapter 5. Graphs (c) and (d) of figure 3.15 show a better performance of the $ChaboNet4k$ model in terms of speed for training and validation. Mean accuracy is also slightly higher: 1.53% of mean accuracy increase is observed with

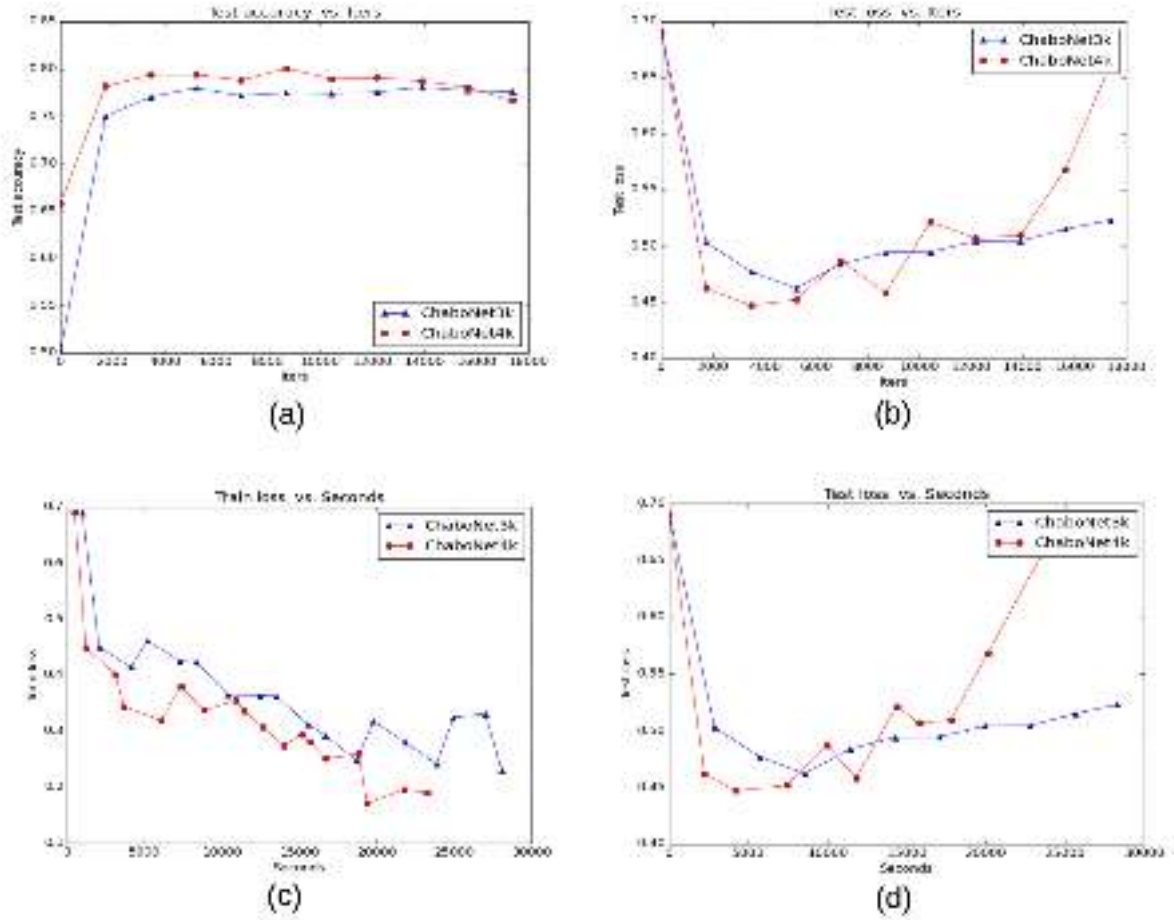


Figure 3.15: Training the network - Accuracy and loss vs iterations and seconds of *ChaboNet3k* and *ChaboNet4k* for "HOLLYWOOD" database : (a) Accuracy vs iterations, (b) Loss on validation data set vs iterations, (c) Train loss vs seconds, (d) Loss on validation data set vs seconds.

merely the same stability of training. The latter is expressed by the standard deviation in the table 3.4.

Table 3.4: The accuracy results on HOLLYWOOD data set

	$ChaboNet3k$	$ChaboNet4k$
$training - time$	$7h47m33s$	$6h27m2s$
$min - Accuracy_{(\#iter)}$	$50.11\%_{(\#0)}$	$65.73\%_{(\#0)}$
$max - Accuracy_{(\#iter)}$	$77.98\%_{(\#5214)}$	$80.05\%_{(\#8690)}$
$avg - Accuracy \pm std$	$77.30\% \pm 0.864$	$78.73\% \pm 0.930$

3.6.3 Validation of the ChaboNet architecture

To evaluate the ChaboNet *architecture* designed for saliency prediction, an experiment was conducted with the HOLLYWOOD dataset. The popular AlexNet [63] and the original LeNet [69] network architectures that are described in section 2.2.2, were used as a base-line with data patches extracted from HOLLYWOOD data. They were trained with two classes in the output corresponding to salient/Non-salient categories of patches.

For AlexNet, the network settings were taken exactly as in [63], that means the same number and size of filters at all layers, the same learning parameteres : leaning rate(0.01), momentum coefficient(0.9), weight decay(0.0005) and number of iterations (450.000). To better visualize, in figure 3.16 the iterations of AlexNet were limited to 70.000. Similarly, the original settings of LeNet were preserved from [69]. Here the number of iterations was 10.000. Chabonet Network training was performed with 17.400 iterations.

Obtained results summarized in figure 3.16 showed that the ChaboNet network outperformed the AlexNet and LeNet architectures (see table 3.5). In fact, with 17.400 iterations, ChaboNet outperformed by 2% in mean accuracy the AlexNet architecture which needed 450.000 iterations. When comparing the 10.000 first iterations of ChaboNet and LeNet, mean accuracy was discovered to be better by more than 20%. Furthermore, the stability of training expressed by small standard deviation is much stronger, see line 4 of the table 3.5.

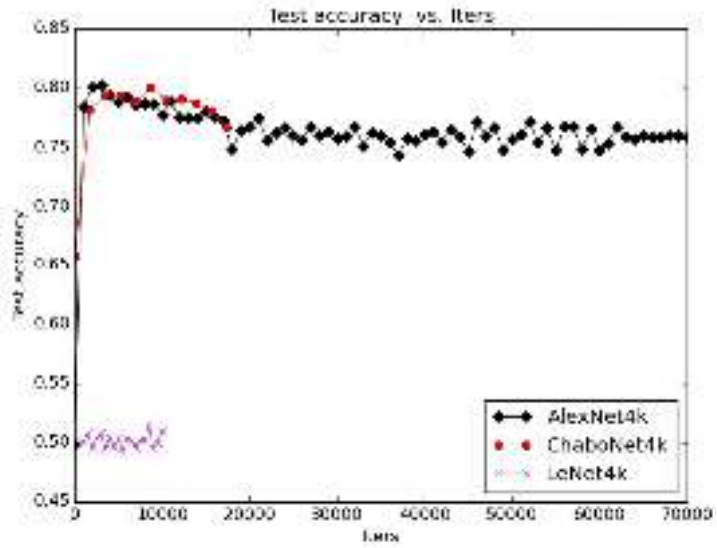


Figure 3.16: Comparison of ChaboNet architecture vs AlexNet and LeNet on Hollywood 4k data set.

Table 3.5: Accuracy results : validation of ChaboNet 4k architecture vs AlexNet and LeNet networks on HOLLYWOOD dataset.

	$ChaboNet4k$	$AlexNet4k$	$LeNet4k$
$min_{(\#iter)}$	$65.73\%_{(\#0)}$	$49,84\%_{(\#0)}$	$49,2\%_{(\#5500)}$
$max_{(\#iter)}$	$80.05\%_{(\#8690)}$	$80,27\%_{(\#3000)}$	$51,56\%_{(\#8500)}$
$avg \pm std$	$78.73\% \pm 0,930$	$76,77\% \pm 6,633$	$50,17\% \pm 0,575$

3.6.4 Evaluation of predicted visual saliency maps

After training and validation of the model on HOLLYWOOD data set, we choose the model obtained at the iteration #8690 having the maximum value of accuracy 80.05%. This model will be used to predict the probability of a local region to be salient. Hence, the final saliency map will be built.

To evaluate our method of saliency prediction, performances were compared with the most popular saliency models from the literature. A spatial saliency models was choosen : Signature Sal [47](the algorithm introduces a simple image descriptor referred to as the image signature, performing better than Itti [52] model), GBVS (regularized spatial saliency model of Harel[41]). and the spatio-temporal model of Seo[113] built upon optical flow.

The comparison of generated predicted saliency maps is performed on the basis of AUC metric (see Chapter 1 for its definition).

In tables 3.6 below, the comparison of Deep CNN prediction of pixel-wise saliency maps with the Gaze Fixations Maps (GFM) is shown.

The quality of predicted maps is compared with prediction by classical saliency models (Signature Sal, GBVS, Seo) also compared to the same reference: GFM. The comparison is given in terms of the widely used AUC metric [67]. Mean value of the metric for each saliency model compared to the GFM is given together with standard deviation for a sample of videos. Hence, in table 3.6 the maps built on HOLLYWOOD database with its best patch saliency prediction model $Chabonet4K$ are compared with GBVS, Signature Sal, Seo.

The best AUC metric values are underscored. It can be stated that in general spatial models (Signature Sal, GBVS or Itti) performed better in half of the tested videos. This is due to the fact that these videos contain very contrasted areas in the video frames, which attract human gaze. They do not contain areas having an interesting residual motion. Nevertheless, the $ChaboNet4K$ model sytematically outperforms Seo's model which uses motion features. This shows definitively that the use of a Deep CNN is a way for prediction of visual saliency in video scenes.

Table 3.6: The comparison of AUC metric of gaze fixations 'GFM' vs predicted saliency 'GBVS', 'SignatureSal' and 'Seo') and our ChaboNet4k for the videos from HOLLYWOOD data set

VideoName	$TotFrame = 2248$	GFM vs GBVS	GFM vs SignatureSal	GFM vs Seo	GFM vs ChaboNet4k
clipTest56	137	$0,76 \pm 0,115$	$0,75 \pm 0,086$	$0,64 \pm 0,116$	$0,77 \pm 0,118$
clipTest105	154	$0,63 \pm 0,169$	$0,57 \pm 0,139$	$0,54 \pm 0,123$	$0,69 \pm 0,186$
clipTest147	154	$0,86 \pm 0,093$	$0,90 \pm 0,065$	$0,70 \pm 0,103$	$0,81 \pm 0,146$
clipTest250	160	$0,74 \pm 0,099$	$0,69 \pm 0,110$	$0,47 \pm 0,101$	$0,71 \pm 0,180$
clipTest350	66	$0,65 \pm 0,166$	$0,68 \pm 0,249$	$0,57 \pm 0,124$	$0,72 \pm 0,177$
clipTest400	200	$0,75 \pm 0,127$	$0,67 \pm 0,110$	$0,60 \pm 0,106$	$0,71 \pm 0,146$
clipTest451	132	$0,70 \pm 0,104$	$0,59 \pm 0,074$	$0,57 \pm 0,068$	$0,63 \pm 0,151$
clipTest500	166	$0,82 \pm 0,138$	$0,84 \pm 0,150$	$0,75 \pm 0,152$	$0,84 \pm 0,156$
clipTest600	200	$0,75 \pm 0,131$	$0,678 \pm 0,149$	$0,53 \pm 0,108$	$0,71 \pm 0,180$
clipTest650	201	$0,72\pm,106$	$0,74 \pm 0,087$	$0,61 \pm 0,092$	$0,70 \pm 0,078$
ClipTest700	262	$0,74 \pm 0,128$	$0,76 \pm 0,099$	$0,50 \pm 0,059$	$0,78 \pm 0,092$
clipTest800	200	$0,70 \pm 0,096$	$0,75 \pm 0,071$	$0,53 \pm 0,097$	$0,66 \pm 0,141$
ClipTest803	102	$0,86 \pm 0,106$	$0,87 \pm 0,068$	$0,73 \pm 0,148$	$0,88 \pm 0,078$
ClipTest849	114	$0,75 \pm 0,155$	$0,91 \pm 0,070$	$0,55 \pm 0,122$	$0,74 \pm 0,132$

Table 3.7, presents the time needed for testing one patch and the creation of the saliency map across one frame with a stride of 5 pixels.

Table 3.7: Time for testing one patch and one frame of video.

	machine $8\mu p$	machine $20\mu p$	machine $2 \times 14 cores \mu p$
patch 100×100	0.015s	0.028s	0.011s
frame 720×576	42.31s	18.49s	8.56s

3.7 Conclusion

Hence, in this chapter, we proposed our solution for prediction of saliency maps in video in the framework of Deep learning. It consists in two steps. First, a deep convolutional network to predict salient areas (patches) in video content was designed. Then dense predicted visual saliency maps was computed on the basis of sparce patch classification results.

We have built an adequate Deep CNN architecture on the basis of Caffe CNN. Deep CNNs being sensitive to noise in training data, we proposed an adapted solution for reducing it. In our case of salincy prediction of image patches, the video production rules

such as the rule of thirds used for Non-salient patches prediction allowed for increase of accuracy.

While the state-of-the art research used only RGB primary values for saliency prediction in visual content, we have shown that for video, adding of features expressing sensitivity of the human visual system to residual motion, is important.

The performances of prediction with Deep CNNs when different kinds of features were ingested by the network, such as color pixel values only, or color values with residual motion- were compared.

We desigend a relatively shallow Deep CNN architecture and have compared it to similar architectures AlexNet and LeNet. It has showed better prediction power in terms of mean accuracy and stability of training phase.

Finally, a method for building pixel-wise saliency maps, using the probability of patches to be salient, was extensively tested against reference spatial and spatio-temporal saliency prediction models.

In the next chapter, we further explore the power of our Deep CNN architecture using research findings on prediction of saliency by classical models.

Chapter 4

Specific saliency features for deep learning

4.1 Introduction

On the contrary to still natural images where saliency is "spatial", based on color contrasts, saturation contrasts, intensity contrasts $\cdots$, the saliency of the video is also based on the motion information of the objects with regard to the background. In the previous chapter, we have briefly introduced motion features that we have added to primary colour values in order to build a specific saliency predictor in video. In this chapter, we will formalize them and go deeper in our experiments. Next in this chapter, we are also interested in combination of learnt features and "engineered" features. Indeed, in the classical saliency models [52], [41], [17], [15], the features were calculated on the basis of phsycho-physiological findings on the sensitivity of HVS to above mentionned contrasts, colours and orientations. These features are then integrated via fusion of feature maps accordingly to the feature integration theory of Treisman and Gelade [125]. Hence the question that we ask in this chapter is "Would known engineered features for saliency prediction improve the prediction accuracy with our designed architecture". Here, we select one model for feature computation and perform a set of experiments on the proposed architecture, integrating all methods of data and feature selection.

4.2 Feature maps

4.2.1 Residual motion feature maps

In video, motion in the frame is a strong visual attractor[95]. Visual attention is not attracted by the motion in general, but by the difference between the global motion in the scene, expressing the camera work, and the "local" motion, that one of a moving object[17] . This difference is called the "residual motion" [87]. In the previous chapter, we have proposed to form a feature map expressing residual motion by its magnitude energy. We used the method developed in[17], [87], [37]. The principle of it consists in computation of residual motion as the difference between raw motion vectors estimated on a pixel-wise basis, i.e. optical flow and a global motion model, expressing camera motion, which is estimated from raw motion vectors. Hence calculation of the residual motion is performed in three steps:

- i) the optical flow estimation, Here we used the optical flow estimator from [78].The method is based on clasical Horn&Schunk formulation [46] of error functional to optimize. Its main improvement compared to Horn&Schunk numerical scheme consists in the use of congugated gradinet method for solving the linear system for the components of a motion vector. In the following we will denote thus estimated motion vectors for each pixel in a video frame by $\overrightarrow{M_c}(x, y)$.

- ii) the estimation of the global motion from optical flow accordingly to the first order complete affine model θ. This model calculates the global or dominant motion by reducing three types of movements of the camera (translations, rotations and zooms), the following equation gives the displacement of the block (dx, dy) of the point (x_0, y_0) to the position (x_1, y_1). We will denot the global motion vectors as $\overrightarrow{M_\theta}(x, y)$.

$$\begin{cases} dx_i = a_1 + a_2(x_1 - x_0) + a_3(x_1 - x_0) \\ dy_i = a_4 + a_5(y_1 - y_0) + a_6(y_1 - y_0) \end{cases} \tag{4.1}$$

Here, $\theta = (a_1, a_2, ..., a_6)^T$,

iii)The residual motion $\overrightarrow{M_r}(x, y)$ expresses the difference between the global motion $\overrightarrow{M_\theta}(x, y)$ and the local motion $\overrightarrow{M_c}((x, y))$. The computation of residual motion is fulfilled according to the equation(4.2):

$$\overrightarrow{M_r}(x, y) = \overrightarrow{M_\theta}(x, y) - \overrightarrow{M_c}(x, y) \tag{4.2}$$

An example of residual motion map is given in Figure 4.1.

(a) Original frame of SRC14: $\#frame30$.

(b) Original frame of SRC14: $\#frame31$.

(c) Absolute value of dx component $\#frame30$.

(d) Absolute value of dy component $\#frame30$.

(e) Normalized energy of movement $\#frame30$.

Figure 4.1: Energy of motion and its components on SRC14 ($\#frame30$) from IRCCyN dataset [16] .

In order to link the strength of motion to the dynamics of each frame, as a final feature map we take the squared norm $L2$ of vectors $\overrightarrow{M_r}(x,y)$ normalised by its maximum in the frame. According to our experience this allows to reduce parasite effectos on contrasted static contours, compared to the real motion of objects. The feature map at a pixel position (x,y) is computed acordingly to the following equation :

$$f^{mot}(I,(x,y)) = \frac{\parallel \overrightarrow{M_r}(x,y) \parallel_2^2}{\max_{(x,y)in\Omega} \parallel \overrightarrow{M_r}(x,y) \parallel_2^2} \tag{4.3}$$

The sensitivity of HVS to motion is selective. Daly [23] proposes a pixel-wise linear model of sensitivity accordingly to the speed of motion. Here, the flat area detection is performed by calculating and thresholding energy gradient. The temporal saliency $S_t(I,(x,y))$ is then deduced by filtering the residual motion by the maximum tracking capacity of the eye. Indeed, the authors [17] reported that the human eye can not follow objects with a speed greater than $80°/s$ [23]. Also the value of the motion saliency acheive its maximum between the speed of $6°/s$ and $30°/s$. Psychovisual filtering proposed by Daly [23] follows the following equation:

$$S_t(I,(x,y)) = \begin{cases} \frac{1}{6}\overrightarrow{M_r}(x,y), & if\ 0 \leq \overrightarrow{M_r}((x,y)) < \overrightarrow{v_1} \\ 1, & if\ \overrightarrow{v_1} \leq \overrightarrow{M_r}(x,y) < \overrightarrow{v_2} \\ -\frac{1}{50}\overrightarrow{M_r}(x,y) + \frac{8}{5}, & if\ \overrightarrow{v_2} \leq \overrightarrow{M_r}(x,y) < \overrightarrow{v}_{max} \\ 0, & if\ \overrightarrow{v}_{max} \leq \overrightarrow{M_r}(x,y) \end{cases} \tag{4.4}$$

where $\overrightarrow{v_1} = 6°/s$, $\overrightarrow{v_2} = 30°/s$ and $\overrightarrow{v}_{max} = 80°/s$.

In our work we use a simplified version, supposing that object motion is in the interval of linearity of Daly's model, which it is not too strong. Hence the energy is a good indicator of interest to a moving object.

The choice of primary spatial features to complete the primary RGB values with "engineered" spatial contrasts can be multiple. Indeed various ways of contrast computation were proposed in [17], [103], [128]. In the present research, we resort to the work in [17] which yields coherent results accordingly to previous sutdies in [15].

To prove the significance of the energy of residual motion, we have conducted an experiment. In this experiment, we compute the AUC metric (well described in the first chapter 1.3.3) between gaze fixation map and the energy of residual motion map. We see that such an experience is obligatory to perform since it will count how many gaze fixation will fall on an area having an interesting residual movement. Here, we used the

most older and popular datasets CRCNS [51] and IRCCyN [16] that are created and benchmarked for the task of saliency prediction in natural videos. These two data sets are deeply described and detailed in next chapter 5.3.1. Used videos of CRCNS data set was illustrated in table 4.2.

Results summerized in table 4.1 and 4.3 show an interesting correspondence between gaze fixation and residual motion map especially for the "gamecube02" video of CRCNS database where we got a 0.56 value of auc metric, and for the "SRC23" video of IRCCyN database where we obtain a very interesting result ($auc = 0.68$). In table 4.3, 8 videos on 12 tested videos give an auc value more than 0.55. Here, we can explain the low value of auc for "SRC02", "SRC07" and "SRC13' videos by that objects in movement are not significant for those scenes. This experience can just encourage us to go further and deeply to prove our fundamental idea of the interest of the integration of residual motion as input to deep CNNs.

Table 4.1: The comparison of AUC metric of gaze fixations 'GFM' vs the energy of ResidualMotion map for 890 frames of CRCNS videos.

VideoName	$TotFrame = 890$	GFM vs ResidualMotion
beverly03	80	0.54 ± 0.119
gamecube02	303	0.56 ± 0.152
monica05	102	0.52 ± 0.110
standard02	86	0.499 ± 0.06
tv-announce01	73	0.472 ± 0.181
tv-news04	82	0.535 ± 0.186
tv-sports04	164	0.500 ± 0.147

Table 4.2: Frames of CRCNS videos.

beverly03	gamecube02	monica05	standard02	tv-announce01	tv-news04	tv-sports04

Table 4.3: The comparison of AUC metric of gaze fixations 'GFM' vs the energy of ResidualMotion map for 456 frames of IRCCyN videos.

VideoName	$TotFrame = 456$	GFM vs ResidualMotion
SRC02	37	0.46 ± 0.025
SRC03	28	0.55 ± 0.112
SRC04	35	0.55 ± 0.191
SRC05	35	0.57 ± 0.148
SRC06	36	0.603 ± 0.156
SRC07	36	0.48 ± 0.028
SRC10	33	0.55 ± 0.086
SRC13	35	0.59 ± 0.147
SRC17	42	0.48 ± 0.071
SRC19	33	0.64 ± 0.078
SRC23	40	0.68 ± 0.094
SRC24	33	0.51 ± 0.045
SRC27	33	0.53 ± 0.074

4.2.2 Primary spatial features

For saliency prediction, the primary spatial features such as simple RGB values are frequently used[117]. Nevertheless, feature integration theory [125] stipulates that HVS is sensetive to specific contrasts: colours, brightness, orientations. Hence, we found interesting to add "engineered" contrast features to the input layer of our network. Would it increase the predictive power of a deep architecture? To answer to our question, we used the contrast features from the saliency model [17] .

The choice of features from [17] is conditioned by their relatively low computational cost and a good performance we have stated. The authors propose seven color contrast descriptors. As the color space 'Hue Saturation Intensity' (HSI) is more appropriate to describe the perception and color interpretation by humans, the descriptors of the spatial saliency are built in this color space. Five of these seven local descriptors depend on the value of the hue, saturation and/or intensity of the pixel. These values are determined for each frame I of a video sequence, from a saturation factor f^{sat} and an intensity factor f^{int}, calculated using the equations (4.5),(4.6):

$$f^{sat}(I,i,j) = \frac{Sat(I,i) + Sat(I,j)}{2} \times (k_{min} + (1 - k_{min}) \cdot Sat(I,i)) \tag{4.5}$$

$$f^{int}(I,i,j) = \frac{Int(I,i) + Int(I,j)}{2} \times (k_{min} + (1 - k_{min}) \cdot Int(I,i)) \tag{4.6}$$

Here $Sat(I,i)$ is the saturation of the pixel i at coordinates (x_i, y_i) and the value at $Sat(I,j)$ is the saturation of the pixel at coordinates (x_j, y_j) adjacent to the pixel i.

$Int(I,i)$ and $Int(I,j)$ are the intensity values respectively. The constant $k_{min} = 0,21$ sets the minimum value for the protection of the interaction of pixel i when the saturation approaches zero [17]. Contrast descriptors are calculated by equations (4.7 ... 4.14):

1. *color contrast*: it is obtained from the two factors of saturation and intensity. This descriptor $X_1(I,i)$ is calculated for each pixel i and its eight connected neighbors j of the frame I, as in equation(4.7):

$$X_1(I,i) = \sum_{j \in \eta_i} f^{sat}(I,i,j) \cdot f^{int}(I,i,j) \tag{4.7}$$

2. *hue contrast*: a hue angle difference on the color wheel can produce a contrast. In other words, this descriptor is related to the pixels having a hue value far from their neighbors (the largest angle difference value is equal to 180°), see equation (4.8):

$$X_2(I,i) = \sum_{j \in \eta_i} f^{sat}(I,i,j) \cdot f^{int}(I,i,j) \cdot \Delta^{hue}(I,i,j) \tag{4.8}$$

The difference in color Δ^{hue} between the pixel i and its neighbors $j = 1 \dots 8$ is calculated accordingly to equations (4.9) and (4.10) :

$$\Delta^{hue} = \begin{cases} \Delta^{\mu}(I,i,j) & if\ \Delta^{\mu}(I,i,j) \leq 0.5 \\ 1 - \Delta^{\mu}(I,i,j) & else \end{cases} \tag{4.9}$$

$$\Delta^{\mu}(I,i,j) = |Hue(I,i) - Hue(I,j)| \tag{4.10}$$

3. *contrast of opponents*: the colors located on the opposite sides of the hue wheel are creating a very high contrast. An important difference in tone level will make the contrast between active color ($hue < 0,5 \simeq 180°$) and passive, more salient. This contribution to the saliency of the pixel i is defined by equation (4.11):

$$\begin{cases} X_3(I,i) = \sum_{j \in \eta_i} f^{sat}(I,i,j) \cdot f^{int}(I,i,j) \cdot \Delta^{hue}(I,i,j) \\ if\ Hue(I,i) < 0.5\ and\ Hue(I,j) \geq 0.5 \end{cases} \tag{4.11}$$

4. *contrast of saturation*: occurs when low and high color saturation regions are close to each other. Highly saturated colors tend to attract visual attention, unless a low saturation region is surrounded by a very saturated area. It is defined by equation (4.12):

$$X_4(I,i) = \sum_{j \in \eta_i} f^{sat}(I,i,j) \cdot f^{int}(I,i,j) \cdot \Delta^{sat}(I,i,j) \tag{4.12}$$

with Δ^{sat} denoting the saturation difference between the pixel i and its 8 neighbors j, see equation (4.13):

$$\Delta^{sat}(I,i,j) = |Sat(I,i) - Sat(I,j)| \tag{4.13}$$

5. *contrast of intensity*: a contrast is visible when dark colors and shiny ones coexist. The bright colors attract visual attention unless a dark region is completely surrounded by highly bright regions. The contrast of intensity is defined by equation (4.14):

$$X_5(I,i) = \sum_{j \in \eta_i} f^{sat}(I,i,j) \cdot f^{int}(I,i,j) \cdot \Delta^{int}(I,i,j) \tag{4.14}$$

where Δ^{int} denotes the difference of intensity between the pixel i and its 8 neighbor j.

$$\Delta^{int}(I,i,j) = |Int(I,i) - Int(I,j)| \tag{4.15}$$

6. *dominance of warm colors*: the warm colors -red, orange and yellow- are visually attractive. These colors ($hue < 0.125 \simeq 45°$) are still visually appealing, although the lack of contrast (hot and cold colors in the area) is observed in the surroundings. This feature is defined by equation (4.16):

$$V_6(I,i) = \begin{cases} Sat(I,i) \cdot Int(I,i) & if\ 0 \leq Hue(I,i) < 0.125 \\ 0 & otherwise \end{cases} \tag{4.16}$$

7. *dominance of brightness and saturation*: highly bright, saturated colors are considered attractive regardless of their hue value. The feature is defined by equation (4.17):

$$V_7(I,i) = Sat(I,i) \cdot Int(I,i) \tag{4.17}$$

The normalization ($V_{1\cdots5}(I,i) = \frac{X_{1\cdots5}}{|\eta_i|}$) of the first five descriptors ($X_{1\cdots5}$) by the number of neighboring pixels ($|\eta_i| = 8$) is performed. In [15], [23] it is reported that mixing a large quantity of different features increases the performance of prediction. This is why it is attractive to mix primary features (1-7) with those which have been used in previous works of saliency prediction[117], that is simple RGB planes of a video frame.

In the follow-up of this chapter we will evaluate performance of our designed ChaboNet

architecture with input layers completed with engineered contrat features, but also consider these choices of input layers in the overall testing framework including filtering of noise in "Non-salient" training patches that we proposed in the previous chapter.

4.2.3 Evaluation of parameters of Deep network

The network was implemented using a powerful graphic card Tesla K40m and processor (2×14 cores). Therefore a sufficiently large amount of patches, 256, was used per iteration, see the *batch_size* parameter in equation (3.5). After a fixed number of training iterations, a model validation step is implemented. At this stage the accuracy of the model at the current iteration is computed. In this section, we put in place and study the influence of input features to accelerate the training of the network. Hence, an increase in network accuracy achieves the training stabilization in a lesser number of iterations and then ensures the complexity reduction. In these two experiments, purely random selection process of Non-salient patches was used in our training dataset.

First experiment. To evaluate our deep network and to prove the importance of the addition of the residual motion map, we pretrained two created models with the same parameter settings and architecture of the network: the first one contained R, G and B, primary pixel values in patches (denoted as $DeepSaliency3k$). The $DeepSaliency4k$ presents the model using RGB and the normalized magnitude of residual motion as input data. In this experiment, we have used a big number of epochs ($epochs = 100.15$) in order to ensure more process of the database and therfore to obtain better trained model. The other parameters of the solver ($base\_lr$: 0.001; max_iter : 174000; lr_policy : "$fixed$" ; $momentum$: 0.9 and $weight\_decay$: $4e-05$; $test\_iter$: 1958 $test\_interval$: 1000) are fixed to run this experiment.

The following figure 4.2 illustrates the variations of the accuracy along iterations of the both models 3k and 4k for the "HOLLYWOOD" database.

Table 4.4: The accuracy results on HOLLYWOOD dataset in the first experiment

		3k_model	4k_model
HOLLYWOOD	$min_{(\#iter)}$	$50.1\%_{(\#0)}$	$49.8\%_{(\#0)}$
	$max_{(\#iter)}$	$74.8\%_{(\#3000)}$	$76.6\%_{(\#3000)}$
	$avg \pm std$	$71.6\% \pm 0.018$	$73.2\% \pm 0.020$

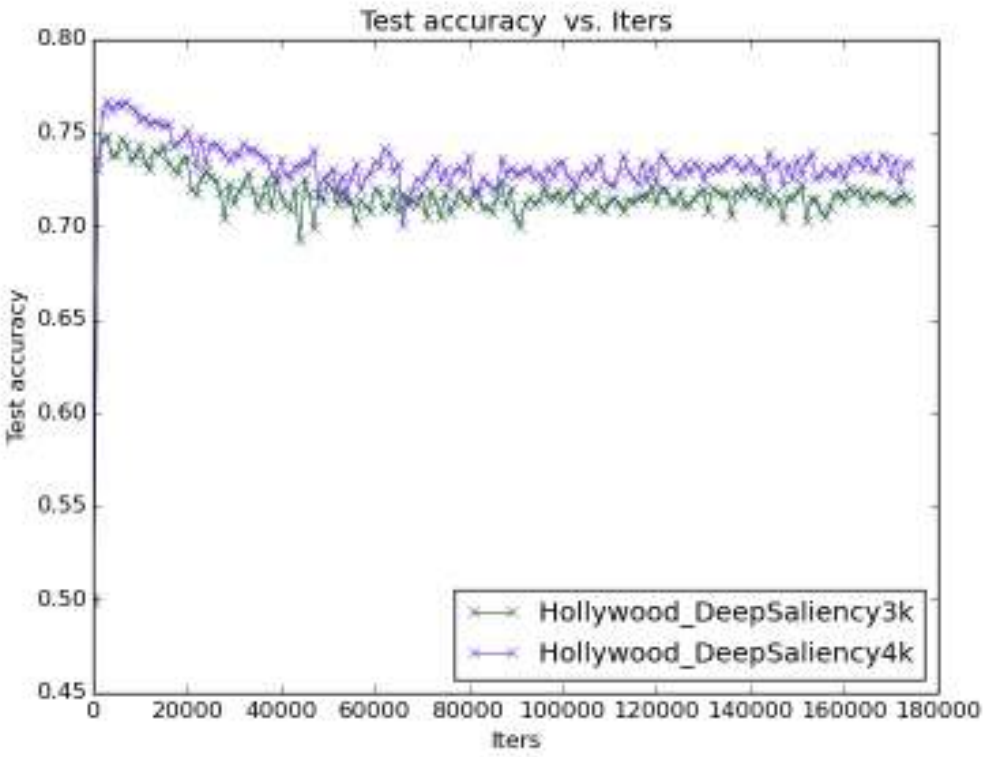


Figure 4.2: First experiment: Accuracy vs iterations of the both models 3k and 4k for "HOLLYWOOD" database.

Second experiment. The second experiment for saliency prediction is conducted when limiting the maximal number of iterations to prevent us from falling into overfitting problem. Instead of increasing the number of training iterations with a limited number of data samples before each validation iteration, as this is the case in the work of [62], we pass all the training set before the validation of the parameters and limit the maximal number of iterations in the whole training process. We used the same equation (3.5) but with a smaller value of epochs in training step ($epochs = 10.15$). Here, a validation step is only started when the whole training data has passed through the network. The equation of validation intervall is written as follows (4.18):

$$Validation\_interval = \frac{\{Total\_images\_number\}_{trainingstep}}{\{batch\_size\}_{trainingstep}} \tag{4.18}$$

In this experiment, the used parameters for the Hollywood dataset are: $test\_iter$: 1958 ; $test\_interval$: 1738; $base\_lr$: 0.001; max_iter : 17400; lr_policy : "$fixed$"; $momentum$: 0.9; $weight\_decay$: $4e-05$; The results are presented in table 4.5 and illustrated in figure 4.3.

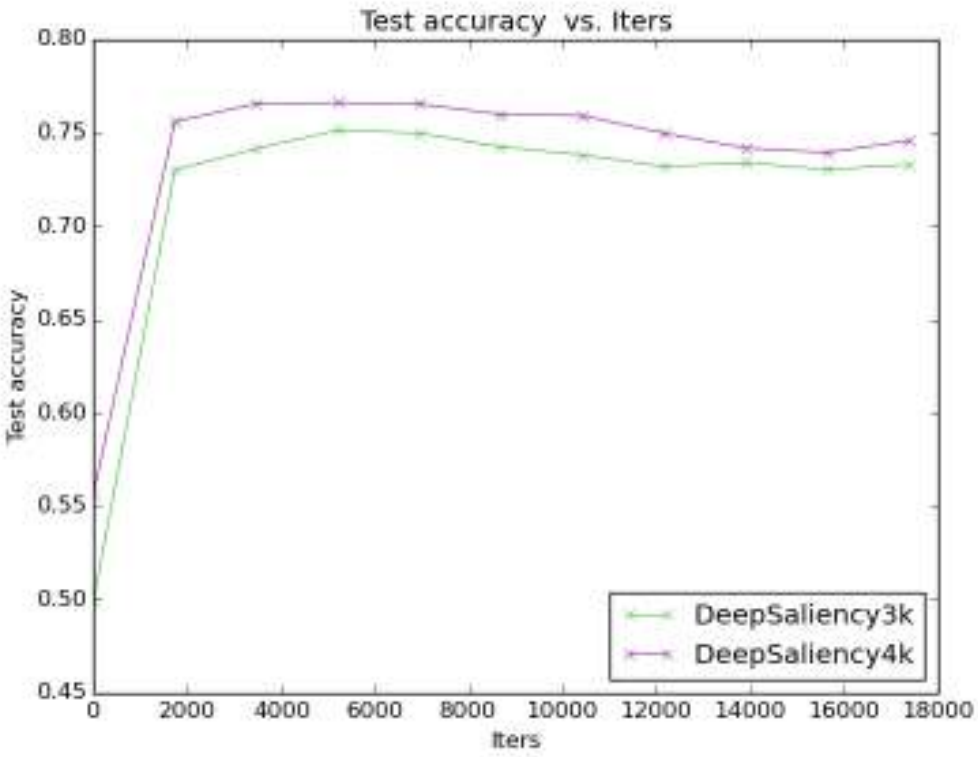


Figure 4.3: Second experiment: Accuracy vs iterations of 3k, 4k for "HOLLYWOOD" database.

Table 4.5: The accuracy results on HOLLYWOOD dataset during the second experiment

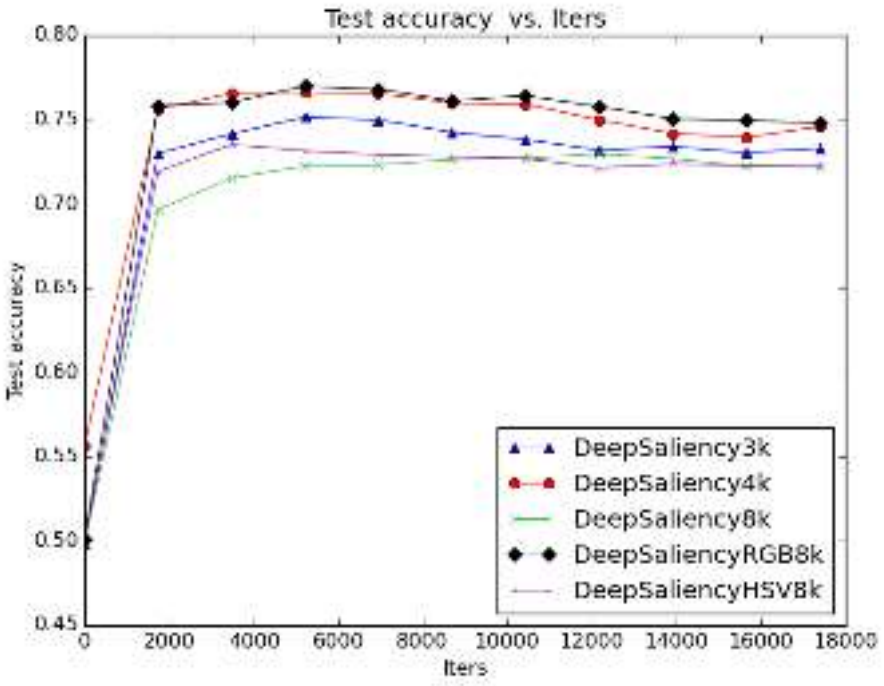
	3k_model	4k_model
$min_{(\#iter)}$	$49.8\%_{(\#0)}$	$55.6\%_{(\#0)}$
$max_{(\#iter)}$	$75.1\%_{(\#5214)}$	$76.6\%_{(\#5214)}$
$avg \pm std$	$71.6\% \pm 0.072$	$73.6\% \pm 0.060$

This drastically decreases (12 times approximately) the training complexity, without the loss of accuracy (see tables 4.4 and 4.5 for 3k and 4k models).

For the HOLLYWOOD database, adding residual motion map improves the accuracy with almost 2% on the 4k model compared to the 3k model. The resulting accuracy of our proposed network along a fixed number of iterations shows the interest of adding the residual motion as a new feature together with spatial feature maps R, G and B. Nevertheless, the essential of accuracy is obtained with purely spatial features (RGB). This is why we add spatial contrast features which have been proposed in classical visual saliency prediction framework [17] in the second experiment in next section.

4.2.4 Evaluation of prediction of saliency of patches

The salient patches were extracted on the basis of Wooding's map according to the process described in section 3.3.1. The maximum number of salient patches extracted by frame was two. In the first experiment we have selected Non-salient patches randomly excluding the area of salient patches. The results of classification accuracy are shown in figure 4.4 for all models, we have considered in our work : $3k - model$ were only RGB values were considered ; $4k - model$ were we added residual motion; $8k - model$ were contrast features together with residual motion ; $RGB8k - model$ where we used all features with RGB values; and finally $HSV8k - model$ presents the HSV values with all features.



(a) Accuracy vs iterations

Figure 4.4: Random selection of Non-salient patches: variations of accuracy along iterations of 3k, 4k, 8k, RGB8k and HSV8k for HOLLYWOOD dataset.

Table 4.6: The accuracy results on HOLLYWOOD dataset during random selection of Non-salient patches experiment.

	3k_model	4k_model	8k_model	RGB8k_model	HSV8k_model
$min_{(\#iter)}$	$49.8\%_{(\#0)}$	$55.6\%_{(\#0)}$	$49.8\%_{(\#0)}$	$50.1\%_{(\#0)}$	$50.1\%_{(\#0)}$
$max_{(\#iter)}$	$75.1\%_{(\#5214)}$	$76.6\%_{(\#5214)}$	$72.9\%_{(\#12166)}$	$76.9\%_{(\#5214)}$	$73.5\%_{(\#3476)}$
$avg \pm std$	$71.6\% \pm 0.072$	$73.6\% \pm 0.060$	$70.1\% \pm 0.067$	$73.5\% \pm 0.078$	$70.5\% \pm 0.068$

We can state that $4k-model$ outperforms all other models in terms of mean accuracy and that adding contrast features does not make improvement as the network learns the contrast features throw its layers. Analyzing the results, we have noticed that purely random selection process of Non-salient patches yielded errors in our training dataset. Hence, we have applied the second method based on 3/3 rule for Non-salient patches selection (see section 3.3.2). The results of this experiment are shown in figure 4.7. We can state that in terms of mean statistics adding "engineered" contrast features to the input layer does not improve prediction accuracy, which remains the best in the case of 4K model.

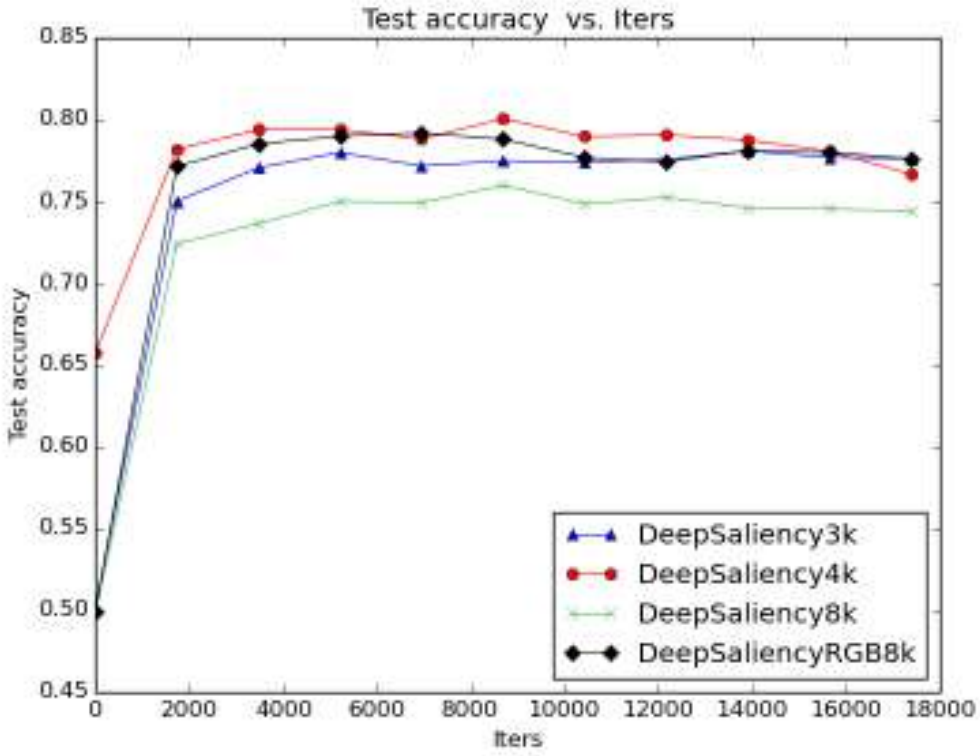


Figure 4.5: Selection of Non-salient patches according to 3/3 rule : Accuracy vs iterations of 3k, 4k, 8k and RGB8k for "HOLLYWOOD" database.

Table 4.7: The accuracy results on HOLLYWOOD dataset during the selection of Non-salient patches according to 3/3 rule.

	3k_model	4k_model	8k_model	RGB8k_model
$min_{(\#iter)}$	$50.11\%_{(\#0)}$	$65.73\%_{(\#0)}$	$49.88\%_{(\#0)}$	$49.92\%_{(\#0)}$
$max_{(\#iter)}$	$77.98\%_{(\#5214)}$	$80.05\%_{(\#8690)}$	$75.98\%_{(\#8690)}$	$79.19\%_{(\#6952)}$
$avg \pm std$	$77.30\% \pm 0.864$	$78.73\% \pm 0.930$	$74.55\% \pm 0.968$	$78.14\% \pm 0.703$

4.2.5 Evaluation of predicted visual saliency maps

In the literature, various evaluation criteria were used to determine the level of similarity between visual attention maps and gaze fixations of subjects like the normalized scanpath saliency 'NSS', Pearson Correlation Coefficient 'PCC', and the area under the ROC curve 'AUC' [85][26]. The "Area under the ROC Curve" measures the precision and accuracy of a system with the goal of categorizing entities into two distinct groups based on their features. The image pixels may belong either to the category of pixels fixated by subjects, either to the category of pixels that are not fixated by any subject. More the area is large, more the curve deviates from the line of the random classifier (area 0.5) and approaches to the ideal bend of the classifier (area 1.00). A value of AUC close to 1 indicates a correspondence between the predicted saliency and the eye positions. While a value close to 0.5 presents a random generation of the salient areas by the model computing the saliency maps. Therefore the objective and subjective saliency differs strongly. In our work, visual saliency being predicted by a deep CNN classifier, we have computed the hybrid AUC metric between predicted saliency maps and gaze-fixations as in [67] (detailed description of AUC metric is given in 1.3.3). The results of the experiments are presented in the tables 4.8, 4.9 and 4.10 below on an arbitrary chosen subset of 12 videos from HOLLYWOOD dataset. The figures depicted in the tables correspond to the maximum value obtained during the training and validation (as presented in table 4.6).

Indeed, with 4k model the results are better for almost all clips, see highlighted figures in table 4.8. For the first experiment the maximal number of iterations was set to 174000 and for the second experiment, this number was fixed 10 times lower. From table 4.8 it can be stated that i) adding primary motion features, such as residual motion improves the quality of predicted visual attention maps whatever is the training of the network. The improvement is systematic and goes up to 38% in case of clipTest105 (in the first experiment); ii) the way to train the network, we propose with lower number of iterations and all training data used does not strongly affect the performances.

From table 4.9 it can be stated that adding primary features to color space improves the quality of predicted visual attention maps. In table 4.9 we compare all our predicted saliency models with gaze fixations. It comes out that more complex models yield better results: up to 42% of improvement in clipTest250. The quality of the prediction of patches (see table 4.6 and figure 4.4) RGB8k_model outperforms HSV8k_model. Therefore, for comparison with reference models from the state of the art, $GBVS$, $SignatureSal$ and spatio-temporal model by Seo [113], named "Seo" we use $RGB8k\_model$, see table 4.10 below.

Table 4.8: The comparison, with AUC metric, of the two experiments for 3K and 4K saliency models vs gaze fixations 'GFM' on a subset of HOLLYWOOD dataset

	First Experiment		Second Experiment	
VideoName	GFM vs 3k_model	GFM vs 4k_model	GFM vs 3k_model	GFM vs 4k_model
clipTest1	$0,58612 \pm 0,19784$	$0,61449 \pm 0,17079$	$0,55641 \pm 0,20651$	$0,77445 \pm 0,14233$
clipTest56	$0,74165 \pm 0,17394$	$0,75911 \pm 0,12509$	$0,65480 \pm 0,19994$	$0,82034 \pm 0,12727$
clipTest105	$0,35626 \pm 0,33049$	$0,74312 \pm 0,19479$	$0,66285 \pm 0,20553$	$0,74740 \pm 0,14689$
ClipTest200	$0,50643 \pm 0,241466$	$0,59407 \pm 0,20188$	$0,53926 \pm 0,21976$	$0,69309 \pm 0,16428$
ClipTest250	$0,548647 \pm 0,240311$	$0,754679 \pm 0,15476$	$0,41965 \pm 0,28409$	$0,72621 \pm 0,15028$
ClipTest300	$0,58236 \pm 0,22632$	$0,66156 \pm 0,16352$	$0,33808 \pm 0,19672$	$0,79186 \pm 0,09732$
ClipTest350	$0,67679 \pm 0,29777$	$0,739803 \pm 0,16859$	$0,47971 \pm 0,40607$	$0,80467 \pm 0,15750$
ClipTest500	$0,58351 \pm 0,20639$	$0,75242 \pm 0,15365$	$0,36761 \pm 0,36777$	$0,82230 \pm 0,15196$
ClipTest704	$0,59292 \pm 0,18421$	$0,68858 \pm 0,16278$	$0,46192 \pm 0,21286$	$0,76831 \pm 0,11186$
ClipTest752	$0,41710 \pm 0,11422$	$0,63240 \pm 0,16870$	$0,25651 \pm 0,25830$	$0,58621 \pm 0,21568$
ClipTest803	$0,67961 \pm 0,24997$	$0,82489 \pm 0,14023$	$0,55019 \pm 0,18646$	$0,87474 \pm 0,06946$
ClipTest849	$0,39952 \pm 0,31980$	$0,67103 \pm 0,20623$	$0,30190 \pm 0,27491$	$0,81148 \pm 0,10363$

Table 4.9: The comparison metric of gaze fixations 'GFM' vs Deep saliency '3k', '4k', '8k' , 'RGB8k' and 'HSV8k' model) for the video from HOLLYWOOD

VideoName	GFM vs 3k_model	GFM vs 4k_model	GFM vs 8k_model	GFM vs RGB8k_model	GFM vs HSV8k_model
clipTest1	$0,55641 \pm 0,20651$	$0,77445 \pm 0,14233$	$0,58518 \pm 0,17991$	$0,725073 \pm 0,168168$	$0,76923 \pm 0,09848$
clipTest56	$0,65480 \pm 0,19994$	$0,82034 \pm 0,12727$	$0,78106 \pm 0,090992$	$0,82244 \pm 0,07295$	$0,81651 \pm 0,06100$
ClipTest105	$0,66285 \pm 0,20553$	$0,74740 \pm 0,14689$	$0,71597 \pm 0,11538$	$0,63652 \pm 0,22207$	$0,81365 \pm 0,08808$
ClipTest200	$0,53926 \pm 0,21976$	$0,69309 \pm 0,16428$	$0,74225 \pm 0,19740$	$0,77948 \pm 0,17523$	$0,68396 \pm 0,17425$
ClipTest250	$0,41965 \pm 0,28409$	$0,72621 \pm 0,15028$	$0,51697 \pm 0,21393$	$0,84299 \pm 0,10787$	$0,69886 \pm 0,13633$
ClipTest300	$0,33808 \pm 0,19672$	$0,79186 \pm 0,09732$	$0,79265 \pm 0,10030$	$0,74878 \pm 0,12161$	$0,83009 \pm 0,08418$
ClipTest350	$0,47971 \pm 0,40607$	$0,80467 \pm 0,15750$	$0,78924 \pm 0,16506$	$0,72284 \pm 0,16996$	$0,80009 \pm 0,232312$
ClipTest500	$0,36761 \pm 0,36777$	$0,82230 \pm 0,15196$	$0,68157 \pm 0,15676$	$0,85621 \pm 0,16137$	$0,88067 \pm 0,09641$
ClipTest704	$0,46192 \pm 0,21286$	$0,76831 \pm 0,11186$	$0,80725 \pm 0,11455$	$0,78256 \pm 0,09523$	$0,79551 \pm 0,071867$
ClipTest752	$0,25651 \pm 0,25830$	$0,58621 \pm 0,21568$	$0,78029 \pm 0,08851$	$0,59356 \pm 0,17804$	$0,76665 \pm 0,07837$
ClipTest803	$0,55019 \pm 0,18646$	$0,87474 \pm 0,06946$	$0,84338 \pm 0,06868$	$0,88170 \pm 0,10827$	$0,85641 \pm 0,06181$
ClipTest849	$0,30190 \pm 0,27491$	$0,81148 \pm 0,10363$	$0,70777 \pm 0,08441$	$0,91089 \pm 0,05217$	$0,71224 \pm 0,07434$

Table 4.10: The comparison of AUC metric gaze fixations 'GFM' vs predicted saliency 'GBVS', 'SignatureSal' and 'Seo') and our RGB8k_model for the videos from HOLLYWOOD dataset

VideoName	GFM vs GBVS	GFM vs SignatureSal	GFM vs Seo	GFM vs RGB8k_model
clipTest1	$0,81627 \pm 0,10087$	$0,69327 \pm 0,13647$	$0,50090 \pm 0,06489$	$0,725073 \pm 0,168168$
clipTest56	$0,76594 \pm 0,11569$	$0,75797 \pm 0,08650$	$0,64172 \pm 0,11630$	$0,82244 \pm 0,07295$
clipTest105	$0,63138 \pm 0,16925$	$0,57462 \pm 0,13967$	$0,54629 \pm 0,12330$	$0,63652 \pm 0,22207$
clipTest200	$0,75904 \pm 0,17022$	$0,87614 \pm 0,10807$	$0,65675 \pm 0,13202$	$0,77948 \pm 0,17523$
clipTest250	$0,74555 \pm 0,09992$	$0,69339 \pm 0,11066$	$0,47032 \pm 0,10193$	$0,84299 \pm 0,10787$
clipTest300	$0,82822 \pm 0,11143$	$0,81271 \pm 0,12922$	$0,75965 \pm 0,13658$	$0,74878 \pm 0,12161$
clipTest350	$0,65136 \pm 0,16637$	$0,68849 \pm 0,249027$	$0,57134 \pm 0,12408$	$0,72284 \pm 0,16996$
clipTest500	$0,82347 \pm 0,13901$	$0,84531 \pm 0,15070$	$0,75748 \pm 0,15382$	$0,85621 \pm 0,16137$
ClipTest704	$0,80168 \pm 0,08349$	$0,85520 \pm 0,06826$	$0,57703 \pm 0,07959$	$0,78256 \pm 0,09523$
ClipTest752	$0,73288 \pm 0,17742$	$0,54861 \pm 0,15555$	$0,71413 \pm 0,13138$	$0,59356 \pm 0,17804$
ClipTest803	$0,86825 \pm 0,106833$	$0,87556 \pm 0,06896$	$0,73847 \pm 0,14879$	$0,88170 \pm 0,10827$
ClipTest849	$0,75279 \pm 0,15518$	$0,91888 \pm 0,07070$	$0,55145 \pm 0,12245$	$0,91089 \pm 0,05217$

Proposed RGB8k_model saliency model turns to be winner more systematically (6/12 clips) than each reference model.

4.2.6 Discussion

Visual saliency prediction with deep CNN is still a recent while intensive research. The major bottle-neck in it is the computation power-and memory requirements. We have shown, that a very large amount of iterations - hundreds of thousands are not needed for prediction of interesting patches in video frames. Indeed, to get better maximal accuracy with smaller amount of iterations we added motion feature, and the maximal number of iterations can be limited (up to 18000 in our case compared to 450000 in AlexNet or 180000 in our first experience) accompanied by another data selection strategy: all data from training set are passed before each validation iteration of the learning, see tables 4.4, 4.5. Next, we have shown that in case of a sufficient training set, adding primary motion features improves prediction accuracy up to 2% in average on a very large data set (HOLLYWOOD test) containing 257733 video frames. Hence the deep CNN captures the sensitivity of Human Visual System to motion.

When applying a supervised learning approach to visual saliency prediction in video, one has to keep in mind that gaze-fixation maps, which serve for selection of training "salient" regions in video frames, not only express the "bottom-up" attention. Humans are attracted by stimuli, but in case of video when understanding a visual scene with time, they focus on the objects of interest, thus reinforcing the "top-down" mechanisms of visual

attention[40]. Hence, the prediction of patches of interest by a supervised learning, we mix all mechanisms: bottom-up and top-down.

In order to re-inforce the bottom-up sensitivity of HVS to contrasts, we completed the input data layers by specific contrast features well studied in classical saliency prediction models. As we could not state the improvement of performance in prediction of saliency of patches in video frames in average (see table 4.5) a more detailed experience clip - by- clip was performed on a sample of clips from HOLLYWOOD dataset when comparing resulting predicted saliency maps. This series of experiments resumed in table 4.11, shows that indeed adding features, expressing local color contrast slightly improves performances with regard to the reference bottom-up spatial (GBVS, SignatureSal) and spatio-temporal models (Seo)). Hence, the mean improvement of AUC of the complete model with motion, contrast features and primary HSV colour pixel values with regard to Itti, Harell and Seo models are 0.00677, 0.01560, 0.15862 respectively. These results are not large (except for Seo model). Hence, we retained the 4k-model definitively for further experiments.

Table 4.11: The mean improvement of the complete model for 1614 frames.

	$\overline{\delta}$(RGB8k_model - GBVS)	$\overline{\delta}$(RGB8k_model - SignatureSal)	$\overline{\delta}$(RGB8k_model - Seo)
AUC	$0,00677 \pm 0,16922$	$0,01560 \pm 0,19025$	$0,15862 \pm 0,21036$

4.3 Conclusion

Hence in this chapter, we completed the RGB pixel values by low-level features of contrast and colour which are easy to compute and have proven efficient in former spatio-temporal predictors of visual attention. Furthermore, we compared different proposed input layers in both frameworks of training data selection: random an dusing the 3/3 rule of visual content production. Despite the accuracy of prediction of saliency of patches is not imporved with added contrast input layer, the quality of predicted saliency maps is sligtly better in terms of AUC metric. What is clearly seen from the experimental results is that adding residual motion maps in the input layer of the network is necessary for prediction of visual saliency in the dynamic video content.

An important point in Deep learning is the availablility of a large amount of training data. Unfortunately in rela-world applications, specifically in health care and medical applications the databases are quite small, and merely encounts hudreds of training samples in various medical studies. Therefore, the next part of the manuscript will be dedicated to studying the transfer learning and its application whena small amount of data are available.

Part III

Transfer Learning

In real life problems such as medical applications, the limited number and size of available data sets could be an obstacle for using powerful Deep learning algorithms. The deep CNNs cannot be trained on a small data. The transfer learning, and specifically a part of it which is "fine tunning" [6] presents a solution to overcome this limits. This part of the manuscript is composed of two chapters. In the first one, the transfer learning scheme for saliency prediction is explained and benchmarked. In the second chapter, the difficulty of the very small amount of data for testing patients with dementia is adressed and a solution is proposed.

Chapter 5

Transfer learning with deep CNN for saliency prediction

5.1 Introduction

The main purpose of transfer learning is to resolve the problem of different data distribution, generally, when the training samples of source domain are different from the training samples of the target domain. Visual saliency models cannot be founded only on bottom-up features, as suggested by feature integration theory. The central bias hypothesis, is not respected neither. In this case, the top-down component of human visual attention becomes prevalent. Visual saliency can be predicted on the basis of seen data.

To predict saliency in video using Deep CNN, the biggest problem is the low number of available video benchmarks with the recorded gaze fixation data. Different databases which have been recorded and made publicly available for e.g. video quality prediction [16] dozens up to one or two hundred of videos. The only public large database is HOLLYWOOD [88] with 1707 videos available with gaze recordings. If saliency prediction in video is realized with a supervised learning approach we are in the framwork of any supervised classification problem requiring sufficient amount of data for training.

In this chapter, the main contribution is to transfer the features learned with the deep network on a large data set in order to train a new network on a small data set with the purpose to predict salient areas.

5.2 Transfer learning with deep networks

The generalization power of Deep CNN classifiers strongly depends on the quantity of data and on the coverage of data space in the training data set. In real-life applications,

e.g. saliency prediction for visual quality assessment [16] the database volumes are small. In order to predict saliency in these small collections of videos, transfer learning approach was nedeed. It presents a technique used in the field of machine learning that increases the accuracy of learning either by using it in different tasks, or in the same task [134]. Transfer learning in Deep CNN presents a powerful tool to enable training on a smaller data set than the base data set [134]. Several studies focused on transferring from higher layers [136], on transferring a pretrained layer to set an unsupervised learning [90].

Transfer learning scheme which we developed in this chapter, is defined as a fine-tuning techniques [6]. Here the authors [6], defined two variants which have been explored in the literature for supervised learning with fine-tuning. The first which was introduced in 2006 in [44] [45] [109] [7], combines two steps: Let x is the raw input, $h_l(x)$ is the ouput of the representation function h at the level l of the input data.

1. Initialize the supervised predictor (parametrized representation function $h_L(x)$ and the linear or non-linear predictor),

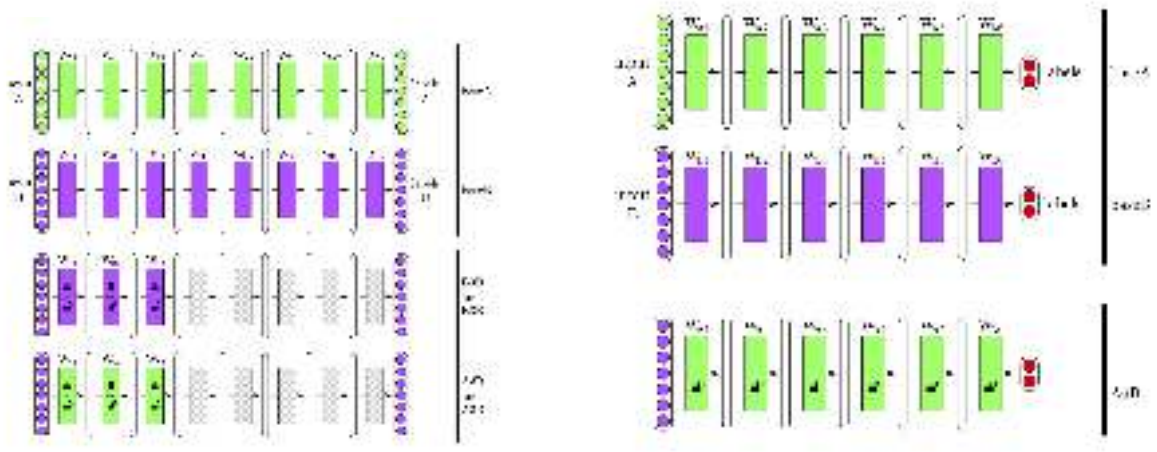
2. Fine-tune the supervised predicor with respect to a supervised training criterion, based on a labled training set of $(x, label)$ pairs, and optimizing the parametres of the supervised predictor.

The second variant of fine-tuning involves using all the levels of representation as input to the predictor. Here, the representation function is fixed and only the linear or non-linear predictor parametres were optimized [72], [73].“ Train a supervised learner taking as input $(h_k(x), h_{k+1}(x), \ldots, h_L(x))$ for some choice of $0 \leq k \leq L$, using a labeled training set of $(x, label)$ pairs.”

Our problem is typically a fine-tunning. We use the same architecture of ChaboNet, the same number and size of filters. Indeed, we propose to adress the same (binary) classification problem on both datasets for prediction of salient or Non-salient class of patches. Its solution consists of two steps: i) learning the whole binary classification model on a large data set, ii) transfer on small data set : initialization of parameters' values in learning process by the optimal parameter values obtained on a large data set. As the classification task is the same in i) and ii) the initialized parameters were supposed to yield a “better” local minimum of loss function, than in the case of a random initialization when training on a small dataset.

In terms of optimization method which is SGD, transfer learning means that the network parameters are not initialized randomly, but their initialization corresponds to a local mimimum of loss function for a large data set. A small database can be considered as different data, thus there won't be the risk of overfitting accordingly to [134]. Starting from

pre-trained parameter values can bring improvement in optimization. Two initialization schemes were tested: that one proposed by Bengio et al. [134] and ours explained in the following (see next figure 5.1).



(a) (b)

Figure 5.1: Comparaison between our proposed scheme of transfer learning and the Bengio's one : a) transfer scheme proposed by Bengio et al. [134] , (b) Our proposed scheme of transfer learning for saliency prediction.

Bengio et al train the models on two datastes: A and B, which he takes as a half of ImageNet database. The coloured rectangles in the figure 5.1 depict network parameters trained independently (see first two rows of the figure a)). Then the two lower rows depict different ways of initialization of parameters in the training proposed by him. In the $selffer$ control, the parameters at the first three layers of the network that is trained on the database B, are used for initialization of training. Parameters of remaining layers are initialized randomly (see B3B and B3B+ notation in the third line of figure 5.1 a). Finally, the network is re-trained. Here, these parameters are either "freezed" , as in B3B scheme or retarined together with the randomly initialized layers (B3B+). Such a scheme makes sense when the two databases are sufficiently large. In case when the database (B) is small the second scheme seems to be more efficient. Here the parameters of three first layers are initialized by the parameters trained on the database A. The argumentation of the authors of using parameters of only first layers for such an initialization [134] consists in saying that the first network layers act merely as wavelet filters, producing features such as blobs and lines on all kinds of databases, and only in deeper layers the network parameters will be adapted to highlight specific structures in images to be classified.

Our classification problem consists in saliency prediction for a given patch in a video

frame. And if we put aside the "top-down" aspects of visual saliency related to the scene interpretation, then the structures related to our classification problems should be the same everywhere accordingly to feature integration theory [125] : contrasts, bright colours, changes in orientations and local singularities of motion. This is why we propose a scheme of A6B+ (see the last line of figure 5.1 b)). This means that the network parameters for training on a small database are initialized for *all* layers by the parameter values optimized on a large database. In the following section we will formalize this model with regard to the parameter optimization method which has been choosen as stochastic gradient descent accordingly to the litterature [62], [117] ...

5.2.1 Stochastic gradient descent 'SGD'

The learning of Deep CNN parameters is frequently done with the technique of Stochastic Gradient Descent (SGD) [12]. The basic iterative equation for finding optimal parameters W optimizing (minimizing in our case) an objective function $f(W)$ is expressed in a classical gradient method [11] [58], by the following equation: *repeat until convergence*

$$\begin{cases} W_{i+1} = W_i - \alpha \frac{\partial}{\partial W_i} f(W_i), \quad i = 1, ...T \\ W_0 = N(0, \sigma^2) \end{cases} \tag{5.1}$$

Here W_i are network parameters of each convolutional layer at the iteration i, T is the total number of iterations, W_0 is the initial value of parameters. The common approach in optimization with Deep NNs consists in a random initialization by a Gaussian distirbution with a zero mean and a small variance of the order of 10^{-3}. We denote it by $N(0, \sigma^2)$, α is the learning rate, and f is the loss function to minimise. The stochastisity in SGD consists in a random selection of packets of data from training set which are used at each iteration. The main SGD problem is that, as a usual gradient descent method, it converges to a local optimum in case when the loss function is not convex. However, it is still the best learning algorithm when the training set is large accordingly to the results reported for visual classification tasks [13].

5.2.2 Transfer learning method

Taking as the basic formulation of the SGD method (see eq. 5.1) and to transfer the classification features obtained from the larger database into the new smaller database as

we have proposed in 2.5, the following principle (5.2) is used for each deep CNN layer.

$$\begin{cases} W_{i+1} = W_i - \alpha \frac{\partial}{\partial W_i} f(W_i) \\ W_0 = W'_n \end{cases} \tag{5.2}$$

with W'_n presents the best learned model parameters pretrained on the large data set.

In optimization for Deep CNNs and namely in Caffe framework [63], more sophisticated method of gradient descent is used, namely the "momentum". Indeed, as stated in [98] in Deep learning, the objective would have the form of a long shallow ravine leading to the optimum and steep walls on the sides. In this case standard SGD will tend to oscillate across the narrow ravine since the negative gradient will point down one of the steep sides rather than along the ravine towards the optimum. The objectives of deep architectures have this form near local optima and thus standard SGD can lead to very slow convergence particularly after the initial steep gains. Momentum is one method for pushing the objective more quickly [54] along the shallow ravine. The momentum update is given by, 5.3. In this equation, we omit any indexes except iteration number i for simplicity:

$$\begin{cases} V_{i+1} = m \cdot V_i - \gamma \cdot \epsilon \cdot W_i - \epsilon \cdot \langle \frac{\partial L}{\partial W} | W_i \rangle_{D_i} \\ W_{i+1} = W_i + V_{i+1} \quad | \quad W_0 = W' \end{cases} \tag{5.3}$$

With $\epsilon = 0.001$- a fixed learning rate, $m = 0.9$ - momentum coefficient, $\gamma = 0.00004$ - wieght decay and W' presents the best learned model parameters pre-trained on the large dataset. The initial value of the velocity V_0 was set to zero. These parameter values are inspired by the values used in [54] with the same fixed learning rate and show the best performances on a large training dataset.

5.3 Experiments and results

5.3.1 Real-life problem : small data sets

To efficiently train a deep network, a very large amount of training data is needed. Hence, in [34] they used hundreds of thousands of windows for training a network in an object recognition task. In the problem of saliency prediction, a very large video data set with available gaze fixations was needed. After reviewing different data sets, we found the only large publicly available data set, the so-called HOLLYWOOD[88], [89]. In this data set

gaze fixations were recorded in a task-driven experiment of action recognition. This data set was used for saliency prediction in video. We have described this dataset in 3.6. Here we just remind the total number of video frames with available gaze fixations: 229825 frames for training and 257733 frames for validation. Application-oriented data sets are usually small.

One of the oldest and well-studied datasets for saliency models benchmarking is CRCNS proposed by Itti [51]. It contains just 46.000 frames. Aonther well-known dataset recorded for video quality assessment tasks is IRCCyN [16] data set. Its number of frames is 61 times smaller than of HOLLYWOOD data set.

We also wish to evaluate the predictive power of proposed Deep CNN classifier in the problem of "top-down" visual attention prediction. From previous research at LaBRI, we have at our disposal an egocentric video dataset with gaze fixations of 31 subjects recorded in a task-driven visual experiment. The subjects were instructed to look at manipulated objects. This dataset, GTEA [27], consists of 17 videos totalling 17632 frames. Therefore, it is also too small for training attention prediction with deep CNNs.
Table 5.1, summarizes the total number of salient and Non-salient patches selected from video frames of the three small data sets.

Table 5.1: Distribution of learning data: total number of salient and Non-salient patches selected from each database.

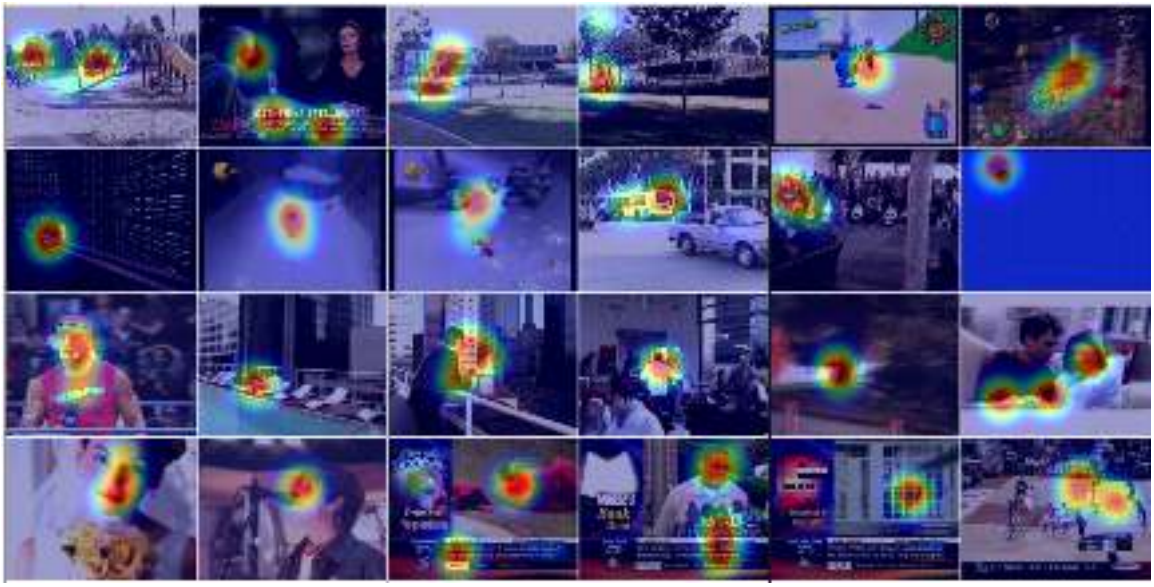
Datasets		training step	validation step
CRCNS	SalientPatch	33370	8373
	Non-salientPatch	30491	7730
	total	63861	16103
IRCCyN-MVT	SalientPatch	2013	511
	Non-salientPatch	1985	506
	total	3998	1017
GTEA	SalientPatch	9961	7604
	Non-salientPatch	9949	7600
	total	19910	15204

In the follow-up of this section we describe these datasets and present statistics of selected training data.

CRCNS data set

In the CRCNS [1] data set [51], 50 videos of 640 × 480 resolution are available with gaze recordings of up to eight different subjects. To create the training, validation and testing set, each video of CRCNS was split according to the following scheme: one frame for testing, one frame for validation and four frames for training set. From the training set, 30370 salient- and 30491 Non-salient patches were selected. From the validation set, a total of 16103 patches were extracted. Table 5.1 resumes the number of salient and Non-salient patches selected for each step : "train" and "validation".

Table 5.2: Preview of CRCNS Data set.



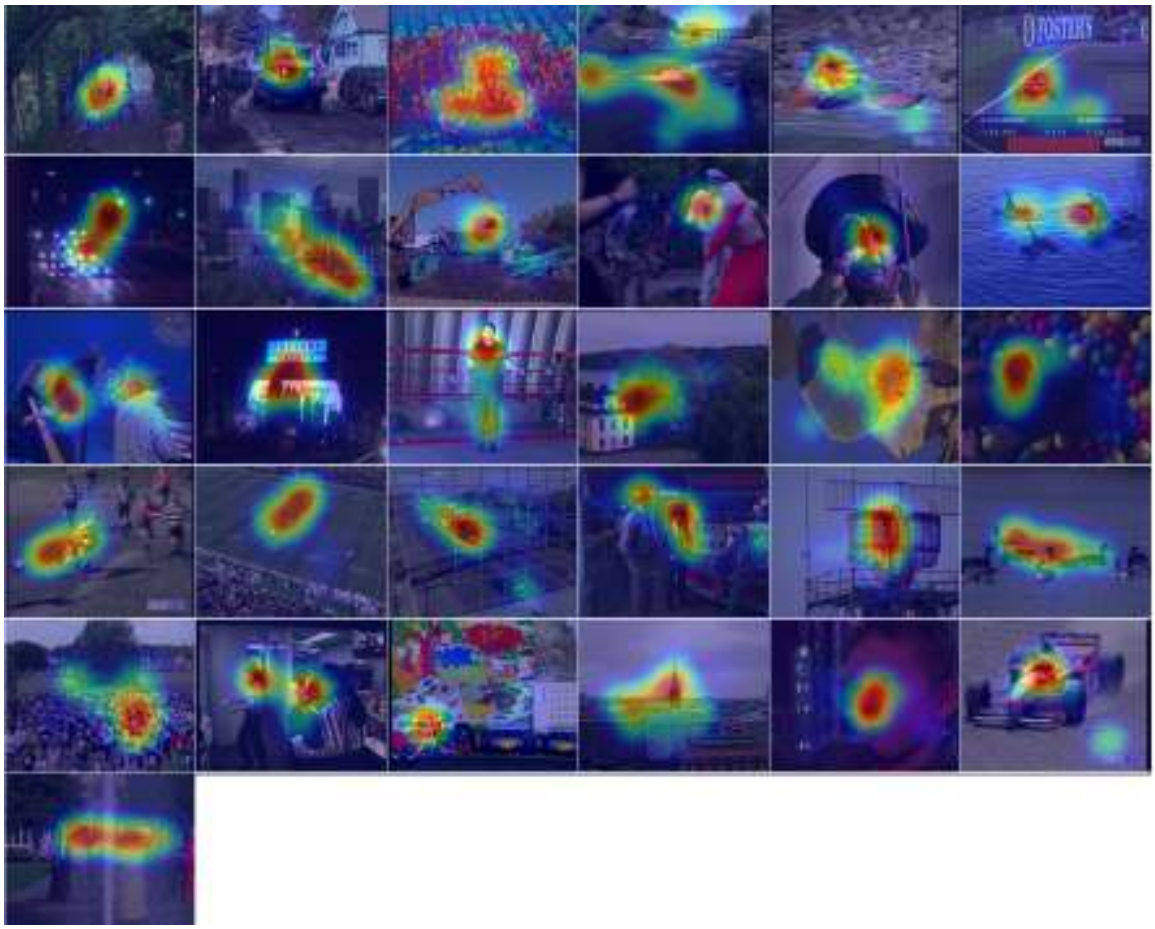
IRCCYN data set

IRCCYN [16] database is composed of 31 SD videos and gaze fixations of 37 subjects. These videos contain certain categories of attention attractors such as high contrast, faces (see table 5.3). However, videos with objects in motion are not frequent. Our purpose of saliency prediction modeling the "smooth pursuit" cannot be evaluated by using all available videos of IRCCyN data set. Videos that do not contain a real object motion were eliminated. Therefore, only SRC02, SRC03, SRC04, SRC05, SRC06, SRC07, SRC10, SRC13, SRC17, SRC19, SRC23, SRC24 and SRC27 were used in experiments, this data set is referenced as IRCCyN-MVT in the following. For each chosen video of this database, one frame is taken for the testing step, one frame for the validation step and four frames for the training step. The distribution of the data between "salient" and "Non-salient"

[1] available at https://crcns.org/data-sets/eye/eye-1

classes is presented in the table 5.1.

Table 5.3: Preview of IRCCyN Data set.

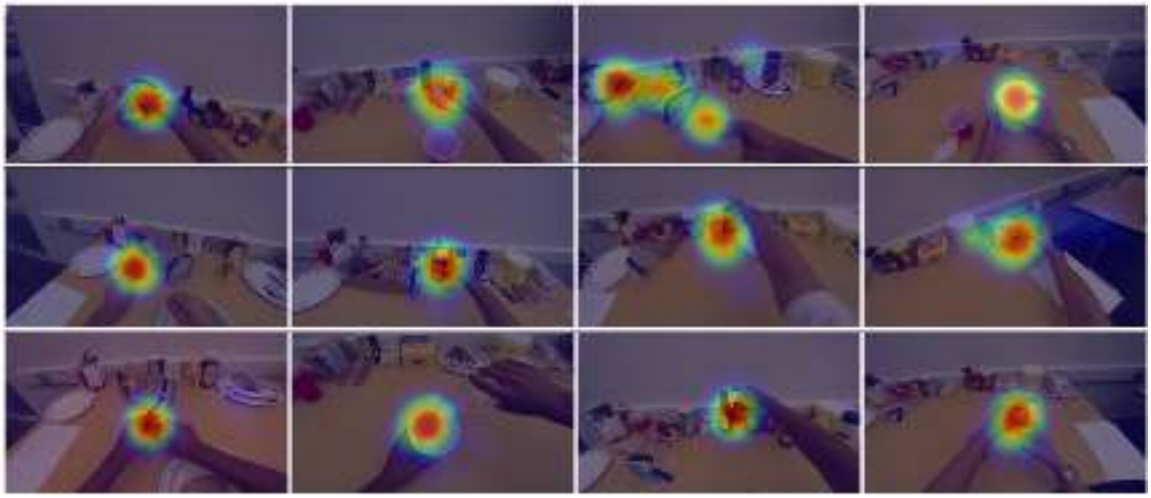


GTEA data set

Egocentric video is becoming popular in various important applications such as monitoring and re-education of patients and disabled persons [56]. Publicly available GTEA corpus [27] contains 17 egocentric videos with a total duration of 19 *min*. GTEA data set consists of videos with $15fps$ rate and a 640×480 pixel resolution. The subjects who recorded the video were preparing meal and manipulating different every day life objects (table 5.4 presents a preview of some video from GTEA corpus). On this data set, we have conducted a psycho-visual experiment with the task of observation of manipulated objects. The gaze fixations have been recorded with a HS-VET 250Hz eye-tracker from Cambridge Research Systems Ltd at a rate of 250 Hz per second. The experiment conditions and the experiment room were compliant with the recommendation ITU-R BT.500-11 [53]. Videos were displayed on a 23 inches LCD monitor with a native resolution of 1920×1080 pixels. To avoid image distortions, videos were not re-sized to screen resolution. A mid-

gray frame was inserted around the displayed video. 31 participants have been gathered for this experiment, 9 women and 22 men. For 3 participants some problems occurred in the eye-tracking recording process. These 3 records were thus excluded. From the 17 available videos of GTEA data set, 10 were selected for the training step with a total number of frames of 10149. And 7 videos with 7840 frames were selected for the validation step. The split of salient and Non-salient patches for the total of 19910 at the training step and 15204 at the validation step is presented in table 5.1 .

Table 5.4: Preview of GTEA Data set.



5.3.2 Learning on small data sets

To apply the proposed transfer learning scheme, the learning of a whole binary classification model on a large data set is required. In chapter 3, the experiment of training and validation of a model for saliency prediction in natural videos was done under the large "HOLLYWOOD" data set. As described in section 3.6 of chapter 3, the best ChaboNet4k model trained on "HOLLYWOOD" data set was obtained at the iteration 8690 with an accuracy value of 80.05%. While the best ChaboNet3k model trained on "HOLLYWOOD" data set was obtained at the iteration 5214 with an accuracy value of 77.98%. These two models were used to initialize features values in learning process of ChaboNet3k and ChaboNet4k on "CRCNS", "IRCCyN-MVT" and "GTEA".

Proposed Transfer learning method on CRCNS data set

Figure 5.2 illustrates the variations of the accuracy and loss along iterations and time in seconds for training $ChaboNet3k$ and $ChaboNet4k$ models on "CRCNS" data set. The

gain of using 4k against 3k as input to the deep CNNs is about 0.22% in terms of mean accuracy. The best model is obtained at the iteration #32500 with an accuracy of 91.66%.

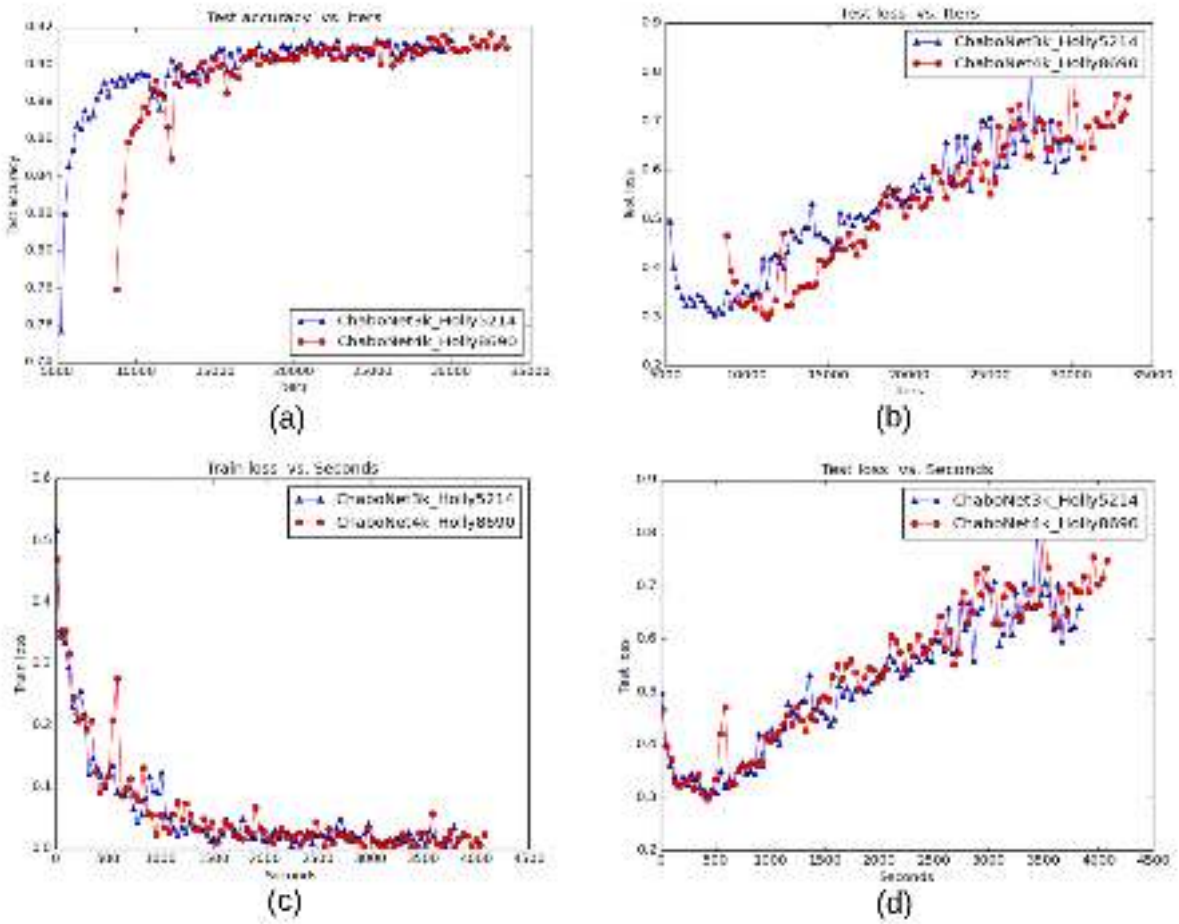


Figure 5.2: Accuracy and loss vs iterations of ChaboNet3k and ChaboNet4k for "CRCNS" database : a) Accuracy vs iterations, (b) Loss on validation data set vs iterations, (c) Train loss vs seconds, (d) Loss on validation data set vs seconds .

Table 5.5: The accuracy results on CRCNS data set

	$ChaboNet3k$	$ChaboNet4k$
$training - time$	$1h3min42s$	$1h7min58s$
$interval - stabilization$	[7500 . . . 30000]	[12500 . . . 33500]
$min - Accuracy_{(\#iter)}$	$87.65\%_{(\#11500)}$	$88.48\%_{(\#15750)}$
$max - Accuracy_{(\#iter)}$	$91.45\%_{(\#28500)}$	$91.66\%_{(\#32500)}$
$avg - Accuracy \pm std$	$90.26\% \pm 0.892$	$90.48\% \pm 0.631$

Proposed Transfer learning method on IRCCyN-MVT data set

Figure 5.3 illustrates the variations of the accuracy along iterations of all models tested for "IRCCyN-MVT". Almost four thousand patches were used for the training of the deep CNN. To overcome the lack of data, the learning was transferred from the best obtained models on "HOLLYWOOD" data set to train the IRCCyN-MVT.

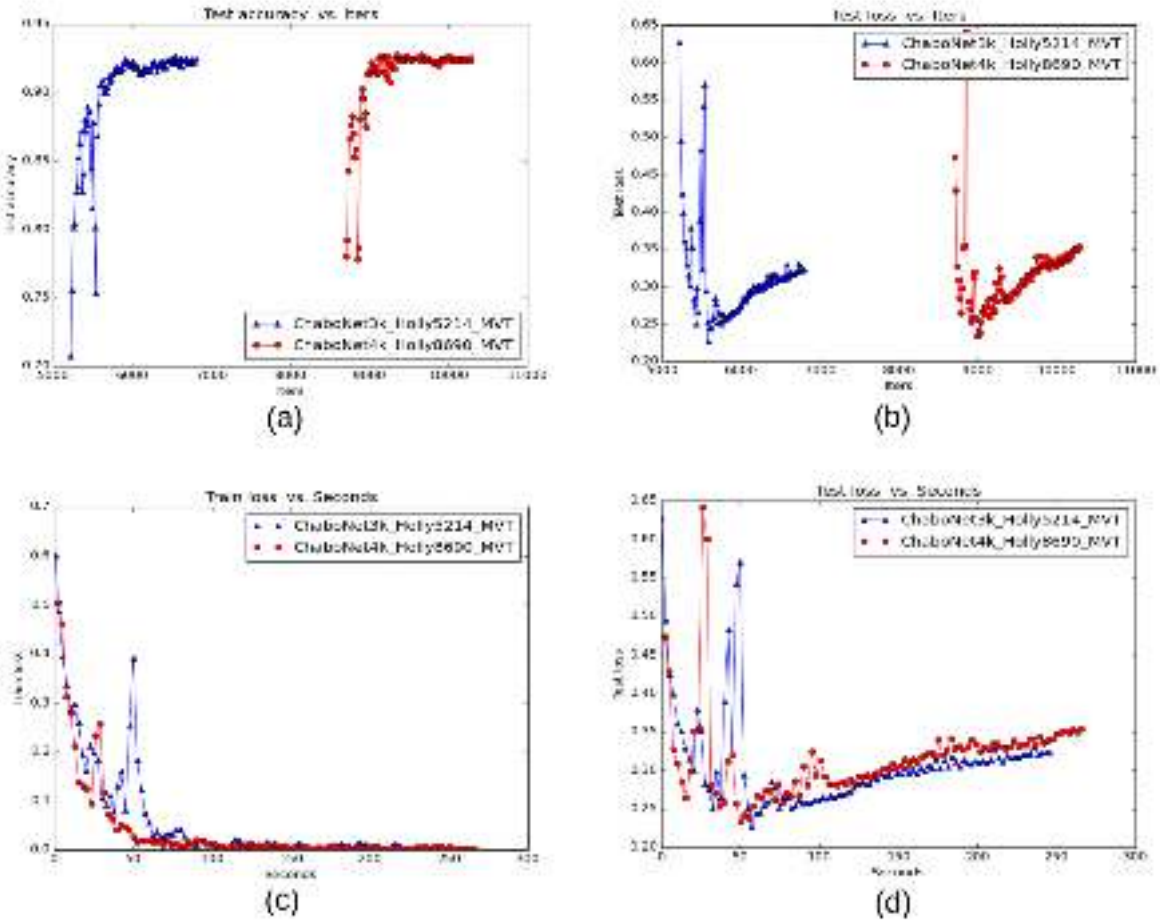


Figure 5.3: Accuracy and loss vs iterations of ChaboNet3k and ChaboNet4k for videos with motion from "IRCCyN-MVT" database : (a) Accuracy vs iterations, (b) Loss on validation data set vs iterations, (c) Train loss vs seconds, (d) Loss on validation data set vs seconds.

With the same number of $epoch = 100$, the $ChaboNet4k$ model reached the interval of stabilization, expressed by smaller standard deviation of accuracy values at all iterations before the $ChaboNet3k$ model did. With an interesting accuracy of 92.77% (see table 5.6) and a small loss of almost 0.35, the best trained $ChaboNet4k$ model was obtained.

Table 5.6: The accuracy results on IRCCyN-MVT data set.

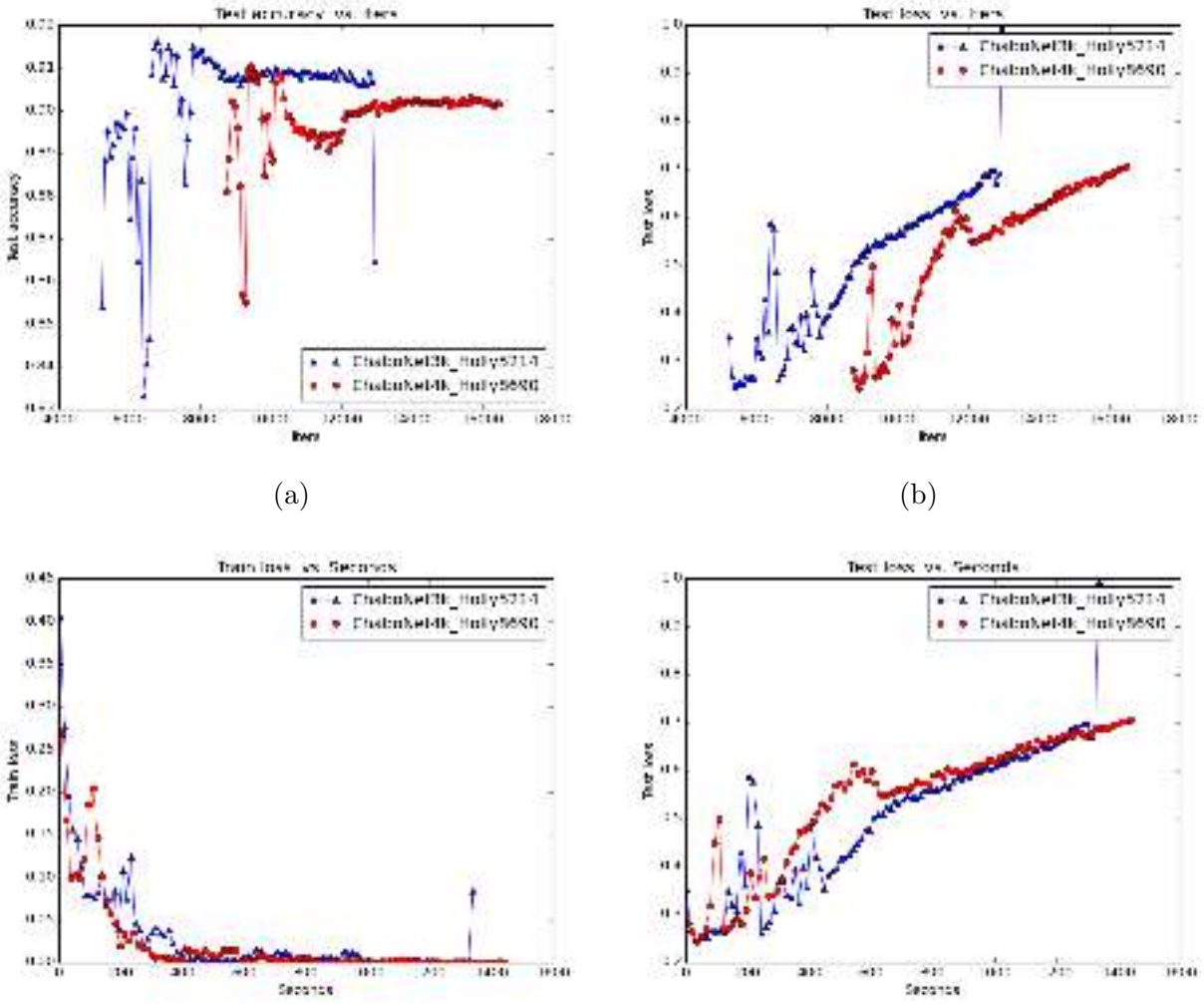
	$ChaboNet3k$	$ChaboNet4k$
$training - time$	$0h4m6s$	$0h4m25s$
$interval - stabilization$	[5584 . . . 6800]	[8976 . . . 10288]
$min - Accuracy_{(\#iter)}$	$89.94\%_{(\#5632)}$	$90.72\%_{(\#9264)}$
$max - Accuracy_{(\#iter)}$	$92.67\%_{(\#6544)}$	$92.77\%_{(\#9664)}$
$avg - Accuracy \pm std$	91.84 ± 0.592	$92.24\% \pm 0.417$

Proposed Transfer learning method on GTEA data set

The results of accuracy on GTEA data set are rather good : average accuracy is about 90% (see table 5.7). Here, $ChaboNet3k$ and $ChaboNet4k$ models were tested. From the plots in figure 5.4, we can see that the $ChaboNet4k$ model is little less efficient than $ChaboNet3k$ model. It is not surprising, the salient patches are predicted by our method according to each visual task : on the Hollywood data set the subjects are instructed to observe actions. They are attracted by the dynamic content of the visual scene. Hence, residual motion is important in the global model. In GTEA data set, the subjects are interested in specific objects be they moving or not. Hence, the spatial appearance is important.

Table 5.7: The accuracy results on GETA data set

	$ChaboNet3k$	$ChaboNet4k$
$training - time$	$0h22m20s$	$0h24min03s$
$interval - stabilization$	[6630 . . . 12948]	[12090 . . . 16458]
$min - Accuracy_{(\#iter)}$	$86,46\%_{(\#7566)}$	$89,80\%_{(\#9750)}$
$max - Accuracy_{(\#iter)}$	$91.61\%_{(\#6786)}$	$90,30\%_{(\#15678)}$
$avg - Accuracy \pm std$	$90.78\%_{(\#0.647)}$	$90,13\%_{(\#0,106)}$



(a) (b)

(c) (d)

Figure 5.4: Accuracy and loss vs iterations of ChaboNet3k and ChaboNet4k for "GTEA" database : a) Accuracy vs iterations, (b) Loss on validation data set vs iterations, (c) Train loss vs seconds, (d) Loss on validation data set vs seconds .

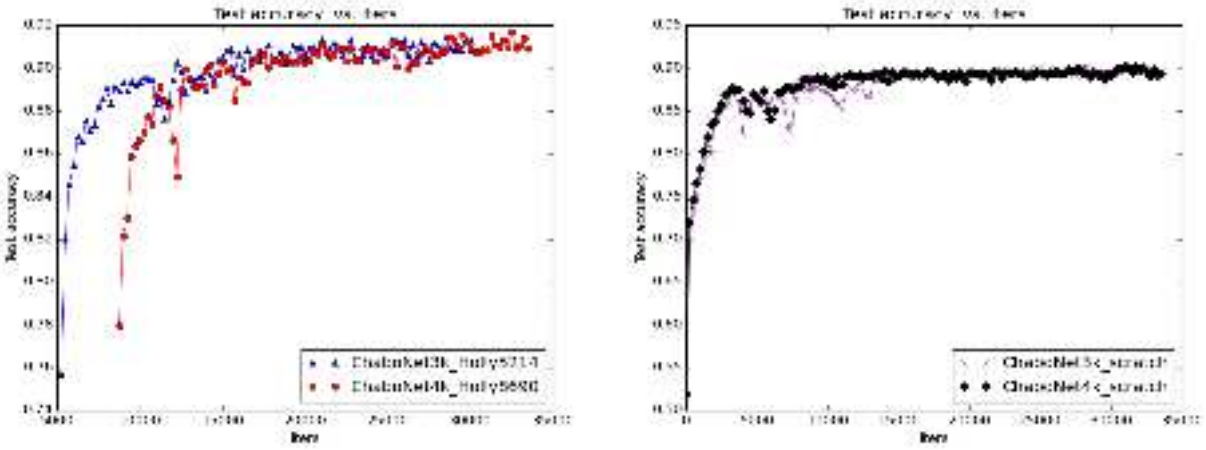
5.3.3 Validation of the proposed transfer learning vs learning from scratch

On three "small" data sets, two experiments were conducted. In the first experiment, the parameters of CNN were initialized randomly from scratch for each layer. In the second experiment the best parameters of the network trained on the large HOLLYWOOD data set were used as the initialization of parameter learning on each "small" data set. The architecture of the CNN remained unchanged in both experiments.

i) First experiment: start training of all ChaboNet layers randomly from scratch.

ii) Second experiment: initialize features parameters of all ChaboNet layers from the best model "features" already trained on the large HOLLYWOOD data set (see section 3.6 of chapter 3) and then fine-tuned on the target data set.

The results presented in figure 5.5, 5.6 and 5.7 show that using the transfer learning of CNN parameters improves not only the value of mean accuracy but also the gain in terms of stabilty of training on the three "small" data sets.



(a) Proposed transfer learning method (b) Learning from scratch

		ChaboNet3k	*ChaboNet4k*
Learning from scratch	$\min_{\#iter}$	$51.72\%_{\#0}$	$52.00\%_{\#0}$
	$\max_{\#iter}$	$90.18\%_{\#28250}$	$90.25\%_{\#31000}$
	avg$\pm std$	$87.11\% \pm 4.655$	$87.85\% \pm 4.169$
Proposed transfer scheme	$\min_{\#iter}$	$75.71\%_{\#5250}$	$77.95\%_{\#8750}$
	$\max_{\#iter}$	$91.45\%_{\#28500}$	$91.66\%_{\#32500}$
	avg$\pm std$	$89.77\% \pm 2.085$	$89.81\% \pm 2.035$

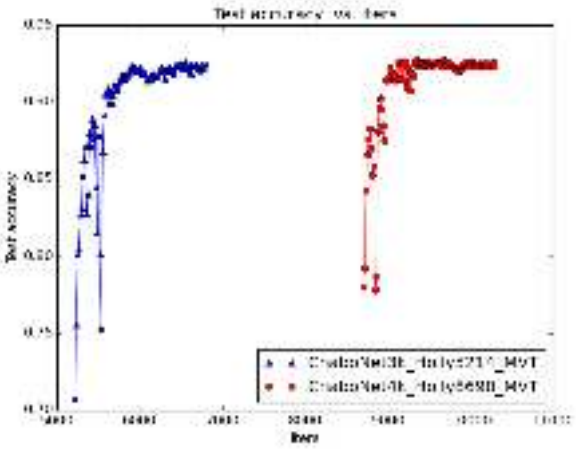
(c) Deep Network performance

Figure 5.5: Evaluation and comparison of our proposed method of transfer learning VS learning from scratch on CRCNS data set.

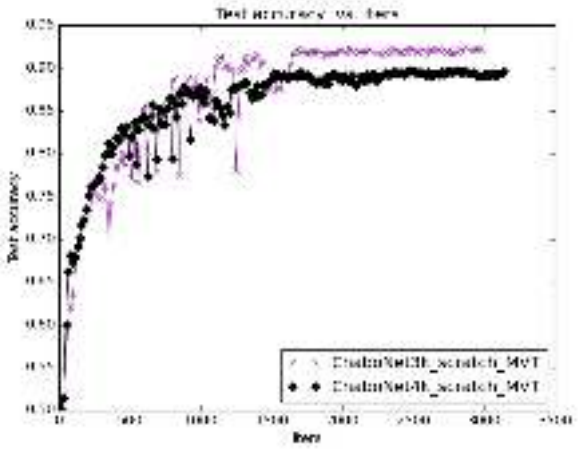
Figure 5.5 illustrates obtained results of the both experiments conducted on CRCNS data set. Here, using transfer learning of the best model trained on HOLLYWOOD data set, we found a higher mean accuracy with almost 2% increase on the both models (figure 5.5). The maximum value of accuracy obtained on the CRCNS data set with the "ChaboNet4k" model is 90.25% at the iteration 31000 using random initialization and 91.66% at the iteration 32500 using pretrained HOLLYWOOD model (see table (c) of figure 5.5). The mean performance of the "ChaboNet4k" model still remains better than performance of the "ChaboNet3k" model.

On the second "small" IRCCyN-MVT database, the following figure 5.6 illustrates

the variations of the accuracy the both models "ChaboNet3k" and "ChaboNet4k" for each experiment. The results show that starting the training with the best parameters of HOLLYWOOD2 model ensures the gain of 6% in the mean accuracy on the "ChaboNet4k" model and the gain of 3% in the mean accuracy on the "ChaboNet3k". The second important point is that with the first experiment of "learning from scratch", the training need more than 1500 iterations to achieve the stabilization in terms of accuracy. Just about twenty iterations is enough to stabilize the accuracy using the proposed transfer learning method.



(a) Proposed transfer learning method



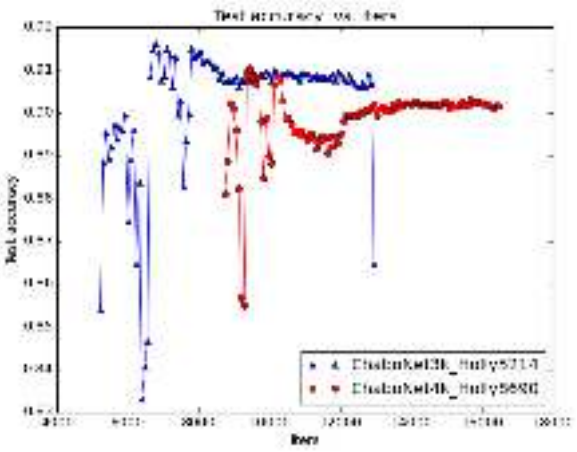
(b) Learning from scratch

		ChaboNet3k	*ChaboNet4k*
Learning from scratch	$\min_{\#iter}$	$50.19\%_{\#0}$	$50\%_{\#16}$
	$\max_{\#iter}$	$92.48\%_{\#2864}$	$89.74\%_{\#2480}$
	avg$\pm std$	$86.46\% \pm 8.592$	$85.40\% \pm 6.818$
Proposed transfer scheme	$\min_{\#iter}$	$70.80\%_{\#5216}$	$77.83\%_{\#8848}$
	$\max_{\#iter}$	$92.67\%_{\#6544}$	$92.77\%_{\#9664}$
	avg$\pm std$	$89.96\% \pm 4.159$	$91.08\% \pm 3.107$

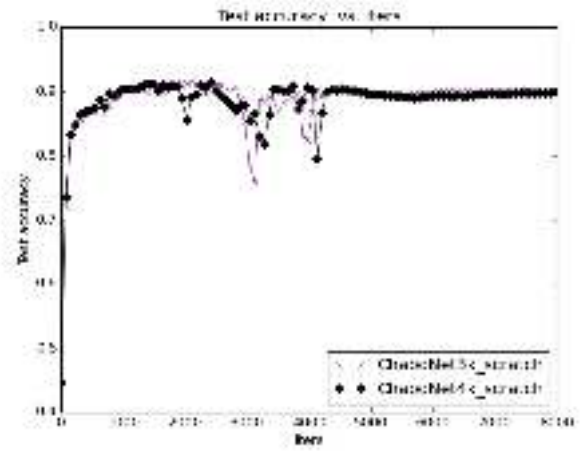
(c) Deep Network performance

Figure 5.6: Evaluation and comparison of our proposed method of transfer learning VS learning from scratch on IRCCyN-MVT data set.

The results of accuracy on "GTEA" data set are rather good : average accuracy is about 90% (see table (c) in figure 5.7). Here, we have tested "ChaboNet3k" and "ChaboNet4k" models. From the plots and the table in figure 5.7, we can see that results are improved in the second experiment with the proposed transfer scheme. Mean accuracy of the both models was executed an increase of almost 2%.



(a) Proposed transfer learning method



(b) Learning from scratch

		ChaboNet3k	*ChaboNet4k*
Learning from scratch	$\min_{\#iter}$	$50.03\%_{\#0}$	$44.63\%_{\#0}$
	$\max_{\#iter}$	$91.45\%_{\#2106}$	$91.50\%_{\#2418}$
	avg$\pm std$	$88.62\% \pm 4.827$	$88.44\% \pm 4.990$
Proposed transfer scheme	$\min_{\#iter}$	$83.32\%_{\#6396}$	$85.48\%_{\#9282}$
	$\max_{\#iter}$	$91.61\%_{\#6786}$	$91.03\%_{\#9438}$
	avg$\pm std$	$90.27\% \pm 1.528$	$89.85\% \pm 0.801$

(c) Deep Network performance

Figure 5.7: Evaluation and comparison of our proposed method of transfer learning VS learning from scratch on GTEA data set.

5.3.4 Validation of the proposed transfer learning vs state-of-the-art transfer learning method

To validate our proposed scheme of transfer learning, the initialization schemes proposed by Bengio et al.[134] was tested. In the research of Bengio et al. [134] addressing object recognition problem, the authors show that the first layers of a Deep CNN learn characteristics similar to the responses of Gabor's filters regardless of the data set or task. Hence in their transfer learning scheme just the three first convolutional layers already trained on a database are used as the initialization of parameters for other database with the same size. The coefficients on deeper layers are left free for optimization, that is initialized from scratch. Here, the context is not the same. Indeed, saliency prediction task is different from object recognition task. Thus the proposal is to initialize all parameters in all layers of the network to train on a small data set by the best model trained on a large data set.

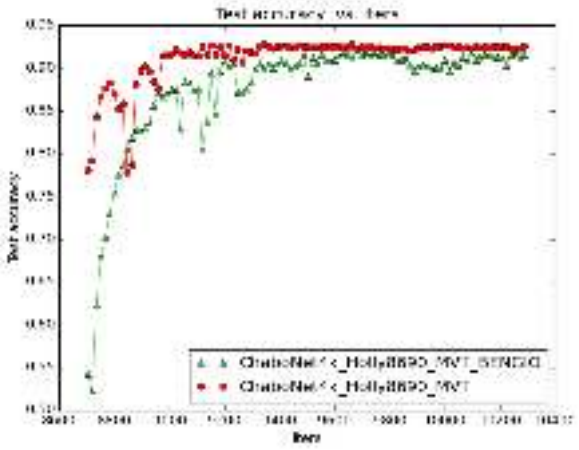
Two experiments were conducted with the same small data set CRCNS and IRCCyN-

MVT , and the same definition of network "ChaboNet4k":

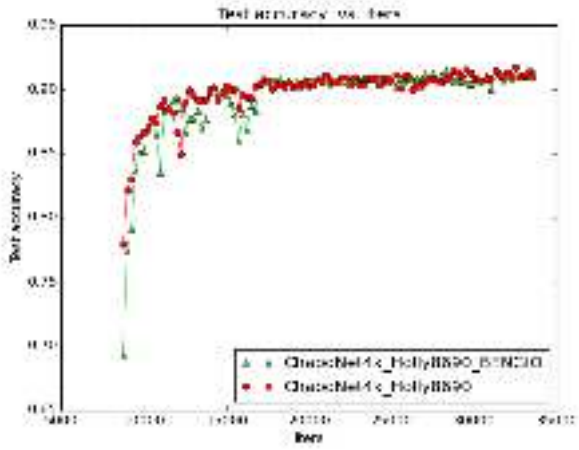
i) Our method: start training of all ChaboNet4k layers from the best model already trained on the large HOLLYWOOD data set (see section 5.2.2).

ii) Bengio's method: the three first convolutional layers are trained on the HOLLYWOOD data set and then fine-tuned on the target data set, other layers are trained on target data set with random initialization.

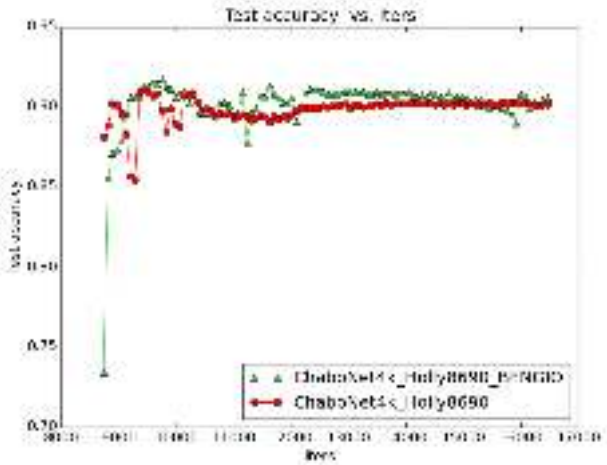
The following figure 5.8 illustrates the variations of the accuracy along iterations of the two methods performed with the data sets "CRCNS" , "IRCCyN-MVT" and "GTEA". One can see less stable behaviour when the transfer method of Bengio et al. is applied.



(a) Comparison on IRCCyN-MVT data set



(b) Comparison on CRCNS data set



(c) Comparison on GTEA data set

Figure 5.8: Evaluation and comparison of our proposed method of transfer learning.

Table 5.8: The accuracy results on IRCCyN-MVT, CRCNS and GTEA dataset.

	Our transfer method			*BENGIO transfer method*		
	IRCCyN-MVT	CRCNS	GTEA	IRCCyN-MVT	CRCNS	GTEA
$max_{(\#iter)}$	$92.77\%_{(\#9664)}$	$91.66\%_{(\#32500)}$	$91\%_{(\#9438)}$	$92.08\%_{(\#9680)}$	$91.55\%_{(\#31250)}$	$91\%_{(\#9750)}$
$avg \pm std$	$91.08\% \pm 3.107$	$89.81\% \pm 2.035$	$89\%_{(\#0.8)}$	$87.48\% \pm 7.243$	$89.37\% \pm 3.099$	$90\%_{(\#1.9)}$

5.3.5 Evaluation of predicted visual saliency maps

After training and validation of the model on CRCNS data set, we choose the model obtained at the iteration #32500 having the maximum value of accuracy 91.66%. This model will be used to predict the probability of a local region to be salient. Hence, the final saliency map will be built. For the IRCCyN-MVT data set, the model obtained at the iteration #9664 with the accuracy of 92.77% is used to predict saliency. In the same manner, the model with the accuracy of 91.03% obtained at the iteration #9438 is used for the GTEA data set.

To evaluate our method of saliency prediction, performances were compared with the most popular saliency models from the literature. Two spatial saliency models were chosen : Itti and Koch spatial model [52], Signature Sal [47] (the algorithm introduces a simple image descriptor referred to as the image signature, performing better than Itti model), GBVS (regularized spatial saliency model of Harel [41]). and the spatio-temporal model of Seo [113] built upon optical flow. The mean value of the AUC metric together with standard deviation were computed.

In tables 5.10 and 5.11 below, we show the comparison of Deep CNN prediction of pixel-wise saliency maps with the gaze fixations "GFM" and compare performances with the most popular saliency prediction models (Signature Sal, GBVS, Seo). Hence, in table 5.9, we compare our *ChaboNet4k* model with the model of Itti, GBVS and Seo.

Table 5.9: The comparison of AUC metric of gaze fixations 'GFM' vs predicted saliency 'GBVS', 'IttiKoch' and 'Seo') and our ChaboNet4k for 890 frames of CRCNS videos

VideoName	$TotFrame = 890$	GFM vs GBVS	GFM vs IttiKoch	GFM vs Seo	GFM vs ChaboNet4k
beverly03	80	0.78 ± 0.151	0.77 ± 0.124	0.66 ± 0.172	0.79 ± 0.118
gamecube02	303	0.73 ± 0.165	0.74 ± 0.180	0.61 ± 0.179	0.82 ± 0.126
monica05	102	0.75 ± 0.183	0.73 ± 0.158	0.54 ± 0.156	0.79 ± 0.133
standard02	86	0.78 ± 0.132	0.72 ± 0.141	0.61 ± 0.169	0.71 ± 0.181
tv-announce01	73	0.60 ± 0.217	0.64 ± 0.203	0.52 ± 0.206	0.63 ± 0.215
tv-news04	82	0.78 ± 0.169	0.79 ± 0.154	0.61 ± 0.162	0.72 ± 0.145
tv-sports04	164	0.68 ± 0.182	0.69 ± 0.162	0.56 ± 0.193	0.78 ± 0.172

Table 5.10: The comparison of AUC metric of gaze fixations 'GFM' vs predicted saliency 'GBVS', 'SignatureSal' and 'Seo') and our ChaboNet4k for the videos from IRCCyN-MVT data set

VideoName	$TotFrame = 1227$	GFM vs GBVS	GFM vs SignatureSal	GFM vs Seo	GFM vs ChaboNet3k	GFM vs ChaboNet4k
src02	37	$0,68 \pm 0,076$	$0,49 \pm 0,083$	$0,44 \pm 0,017$	$0,012 \pm 0,077$	$0,48 \pm 0,073$
src03	28	$0,82 \pm 0,088$	$0,87 \pm 0,057$	$0,76 \pm 0,091$	0.00 ± 0.000	$0,70 \pm 0,149$
src04	35	$0,79 \pm 0,058$	$0,81 \pm 0,029$	$0,59 \pm 0,057$	$0,12 \pm 0,214$	$0,57 \pm 0,135$
src05	35	$0,73 \pm 0,101$	$0,67 \pm 0,122$	$0,48 \pm 0,071$	$0,39 \pm 0,186$	$0,53 \pm 0,128$
src06	36	$0,85 \pm 0,080$	$0,71 \pm 0,151$	$0,73 \pm 0,148$	0.00 ± 0.000	$0,60 \pm 0,180$
src07	36	$0,72 \pm 0,070$	$0,73 \pm 0,060$	$0,57 \pm 0,060$	$0,34 \pm 0,284$	$0,55 \pm 0,135$
src10	33	$0,87 \pm 0,048$	$0,92 \pm 0,043$	$0,82 \pm 0,101$	0.00 ± 0.000	$0,60 \pm 0,173$
src13	35	$0,79 \pm 0,103$	$0,75 \pm 0,111$	$0,64 \pm 0,144$	$0,36 \pm 0,201$	$0,52 \pm 0,138$
src17	42	$0,55 \pm 0,092$	$0,33 \pm 0,099$	$0,45 \pm 0,033$	0.00 ± 0.000	$0,51 \pm 0,098$
src19	33	$0,76 \pm 0,094$	$0,68 \pm 0,086$	$0,59 \pm 0,117$	$0,46 \pm 0,075$	$0,75 \pm 0,123$
src23	40	$0,76 \pm 0,050$	$0,69 \pm 0,070$	$0,58 \pm 0,067$	$0,03 \pm 0,169$	$0,66 \pm 0,105$
src24	33	$0,63 \pm 0,071$	$0,58 \pm 0,054$	$0,55 \pm 0,059$	$0,23 \pm 0,252$	$0,50 \pm 0,052$
src27	33	$0,59 \pm 0,117$	$0,64 \pm 0,091$	$0,52 \pm 0,057$	0.00 ± 0.000	$0,54 \pm 0,106$

Table 5.11: The comparison of AUC metric gaze fixations 'GFM' vs predicted saliency 'GBVS', 'SignatureSal' and 'Seo') and our $4k\_model$ for the videos from GTEA data set

VideoName	$TotFrame = 7693$	GFM vs GBVS	GFM vs SignatureSal	GFM vs Seo	GFM vs ChaboNet4k
$S1\_CofHoney\_C1\_undist$	1099	$0,811 \pm 0,109$	$0,800 \pm 0,091$	$0,578 \pm 0,120$	$0,732 \pm 0,157$
$S1\_Pealate\_C1\_undist$	1199	$0,824 \pm 0,099$	$0,846 \pm 0,080$	$0,594 \pm 0,139$	$0,568 \pm 0,185$
$S1\_Tea_C1\_undist$	1799	$0,770 \pm 0,127$	$0,816 \pm 0,074$	$0,567 \pm 0,135$	$0,745 \pm 0,211$
$S2\_Cheese\_C1\_undist$	499	$0,813 \pm 0,116$	$0,766 \pm 0,0138$	$0,552 \pm 0,127$	$0,643 \pm 0,218$
$S2\_Coffee\_C1\_undist$	1599	$0,802 \pm 0,098$	$0,720 \pm 0,094$	$0,594 \pm 0,116$	$0,636 \pm 0,193$
$S3\_Hotdog\_C1\_undist$	699	$0,768 \pm 0,103$	$0,851 \pm 0,088$	$0,585 \pm 0,114$	$0,415 \pm 0,145$
$S3\_Peanut\_C1\_undist$	799	$0,757 \pm 0,115$	$0,758 \pm 0,135$	$0,519 \pm 0,100$	$0,570 \pm 0,162$

In general, it can be stated from the results on CRCNS data set (table 5.9) that spatial models (Signature Sal, GBVS or Itti) performed better in three tested videos. This is due to the fact that these videos contain very contrasted areas in the video frames, which attract human gaze. They do not contain areas having an interesting residual motion. Nevertheless, the $ChaboNet4K$ model outperforms the Seo model which uses motion features such as optical flow.

However, for IRCCyN-MVT data set, see table 5.10, despite videos without any motion were set aside, the gain in the proposed model is not very clear due to the complexity of these visual scenes, such as presence of strong contrasts and faces.

The comparison for some videos of GTEA data set with different manipulated objects was conducted. In general we can state that spatial models perform better (Signature Sal, GBVS). Nevertheless, our "ChaboNet4k" model outperforms that one of Seo in 4 cases on this 7 examples. This shows that definitly the use of a Deep CNN is a way for prediction of top-down visual saliency in video scenes.

5.4 Conclusion

The transfer learning in the task of saliency prediction is interesting and allows to solve the problem of the insufficiency training data. The transfer learning scheme introduced and applied to the prediction of saliency on small data sets by fine-tuning parameters pre-trained on a large data set (Hollywood) successfully outperforms the state-of-the-art, i.e. Bengio's method.

Hence in this chapter we tackled the problem of prediction of visual attention on video content in a realistic context, when the volume of training data is small. We have developed a transfer learning/fine-tunning approach where the parameters at all layers of the network were initialized by pre-trained on a large data base values. It gives a relatively small, but still a gain compared to the state-of-the art method. Furthermore, the stability of training caracterized by the standard deviation of accuracy along iterations is improved by almost 50%.

The next chapter deals with the second use case of this work which is the application for testing of patients with neuro-degenerative diseases.

Chapter 6

Application of saliency prediction for testing of patients with neuro - degenerative diseases

6.1 Introduction

Studies of visual attention of patients with Dementia such as Parkinson's Disease Dementia and Alzheimer Disease is a promising way for non-invasive diagnostics. Past research showed, that people suffering from dementia are not reactive with regard to degradations on still images [22]. Attempts are being made to study their visual attention relatively to the natural video content [126]. If a degraded visual content is displayed for patients with dementia, the delays in their reactions on novelty and "unusual" novelty of the visual scene are expected. Nevertheless, large-scale screening of population is possible only if sufficiently robust automatic prediction models can be built. In the medical protocols the detection of Dementia behavior in visual content observation is always performed in comparison with healthy, "normal control" subjects. Hence, it is a research question per see as to develop an automatic prediction models for specific visual content to use in psycho-visual experience involving Patients with Dementia (PwD). The difficulty of such a prediction resides in a very small amount of training data both in terms of quantity subjects as in terms of quantity of specifically post-produced content. In litterature, the difference in saccadic eye movements of PwD compared to control subjects of the same age have been stated [22]. We hypothesize that a difference in visual fixation maps of healthy subjects and PwD will also exist.

In this chapter we aim to build an automatic prediction model of attention of healthy subjects with regard to intentionally degraded content. Hence, the first study conducted

in this framework aimed to identify the difference of reaction of healthy subjects to "normal" dynamic video content and "unusual distractors", which are intentionnaly introduced degaradation. Then, taking into account a small amount of specifically produced video content, optimal transfer learning strategy for training the model in case of very small amount of training data was deployed. The comparison with gaze fixation maps and classical visual attention prediction models was performed. Results are interesting regarding the reaction of normal control subjects against degraded areas in videos.

6.2 Material and methods

To analyze the anomalies of eye saccades at the prodromal stage of neurodegenerative diseases, and respecting the bio-medical research protocol "LYLO" [123] cf, Appendix, two types of full HD video (1920x1080 with the frame-rate of 25 frames per second) were created : normal videos and a set of artificially degraded videos. The purpose is to conduct a psycho-visual experiment to compare fixations on the degraded regions and the induced visual attention maps for normal control subjects and patients with dementia. In this experiment, medical researchers choose the nature and locus of degradations as texture modification. The duration of video clips was chosen to avoid the phenomena of visual fatigue and is of 28 seconds. Hence, this database is specified by a very small size: only 700 frames in each of the two video clips. The original "normal" video content and degraded one were displayed to the normal control subjects and are supposed to be displayed to PwD in a free viewing conditions. We will now describe the nature and the methods for creating degradations on natural video for this purpose.

6.2.1 Definition of degradations

As the original material for creation of "degraded videos" we have selected full HD video (1920 x 1080 pixels) at 25 fps produced in the framework of the project ICOS-HD at Labri and available on OpenVideo.org platform [93], [8]. Each natural video was processed frame by frame in order to create naturally degraded areas.

Hence, degradations were added, such as Gaussian blur or pixelation, on objects in specific areas at different locations in the video frames maintaining spatial coherency along the time (the objects in the center, right or left, top or bottom). Two kinds of placement were performed: i) on an environmental object, ii) on the background. We avoided placing the degradations on moving objects in the video sequences, as they are natural attractors of attention and the goal of the experiment is to measure the curiosity of the subjects with regard to an unusual content. A Gaussian blur was used with the

size of [30x30]. The spread parameter value was chosen as $\sigma = 50$ accordingly to the size of degraded areas with regard to the resolution of frames (see Gaussian equation 6.1).

$$G(x, y) = \frac{1}{2\pi\sigma^2} e^{\frac{-(x^2+y^2)}{2\sigma^2}} \tag{6.1}$$

After applying the degradation on videos, two video sequences have been created, one for "normal" video and one for degraded videos (see figure 6.1). Hence, each sequence is composed of two videos separated by a black screen of 200 milliseconds duration ensuring resetting the status of the visual attention of observers. The overall duration of thus

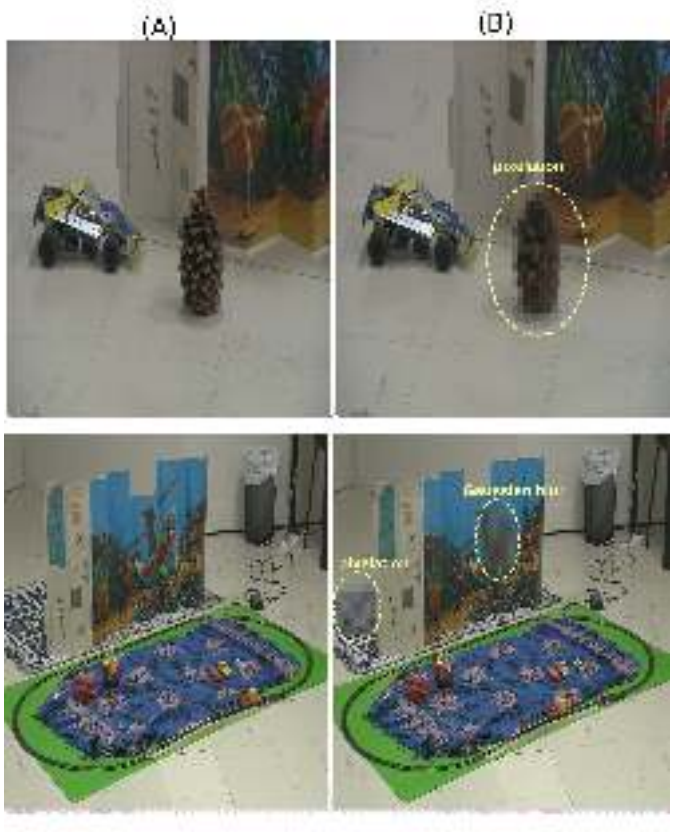


Figure 6.1: (A) Normal video, (B) degraded video.

produced video clips were 56s corresponding to the amount of 1411 frames.

6.2.2 Validation of degraded maps: Creation of visual attention maps

To validate the degraded sequences, visual attention maps of normal control subjects were compared on the two kinds of videos : normal sequences first and then the degraded ones.

In the following the psycho-visual experiment for recording the fixations and the reference method for creation of visual attention maps are described.

Setting up of the psycho-visual experiment

The human visual attention is measured by recording the movement of the eye. Eye movements are portrayed with a sequence of saccades, fixations and smooth pursuits. The jerks are movements with large amplitude that allow exploration of the visual field. Instead, the bindings are micro-saccades with a low amplitude that place the object of interest on the fovea. Consequently, fine details are extracted over the fasteners. Smooth pursuits are triggered when tracking a moving object [39]. Their role is to keep the object on the fovea.

Eye-trackers are used to record and measure eye movements. These devices emit an infrared light and contain an infrared camera. Infrared light illuminates the eye and the camera records its movement. The recording of eye movements represents a digital processing that is required to follow the white spot and black pupil (see figure 6.2).

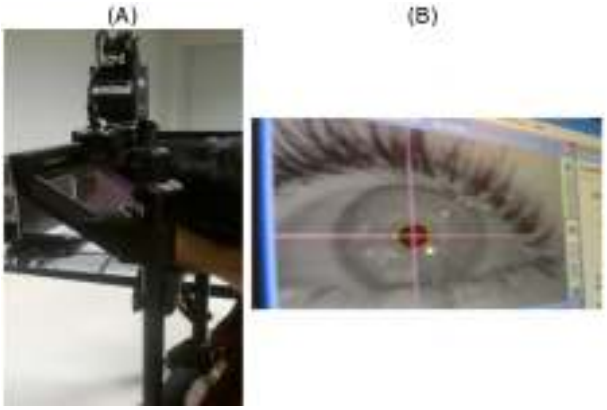


Figure 6.2: Digital recording of the eye movement (A) Eye tracker provides an infrared mirror reflecting infrared light. (B) The benchmark for measuring eye movements (the white spot on the pupil presents a reflection of the infrared light on the eye).

During the psycho-visual experiment and respecting the LYLO protocol [123], recorded eye movements were obtained with the Cambridge Technology EyeTracker device (see figure 6.3). It contains a monocular infra-red camera and ensures recording frequency of 250 Hz.

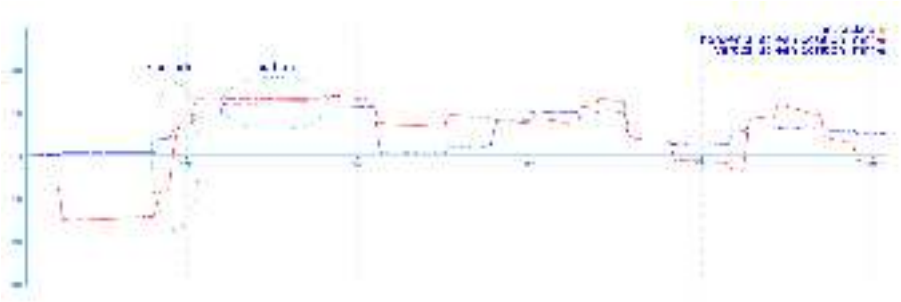
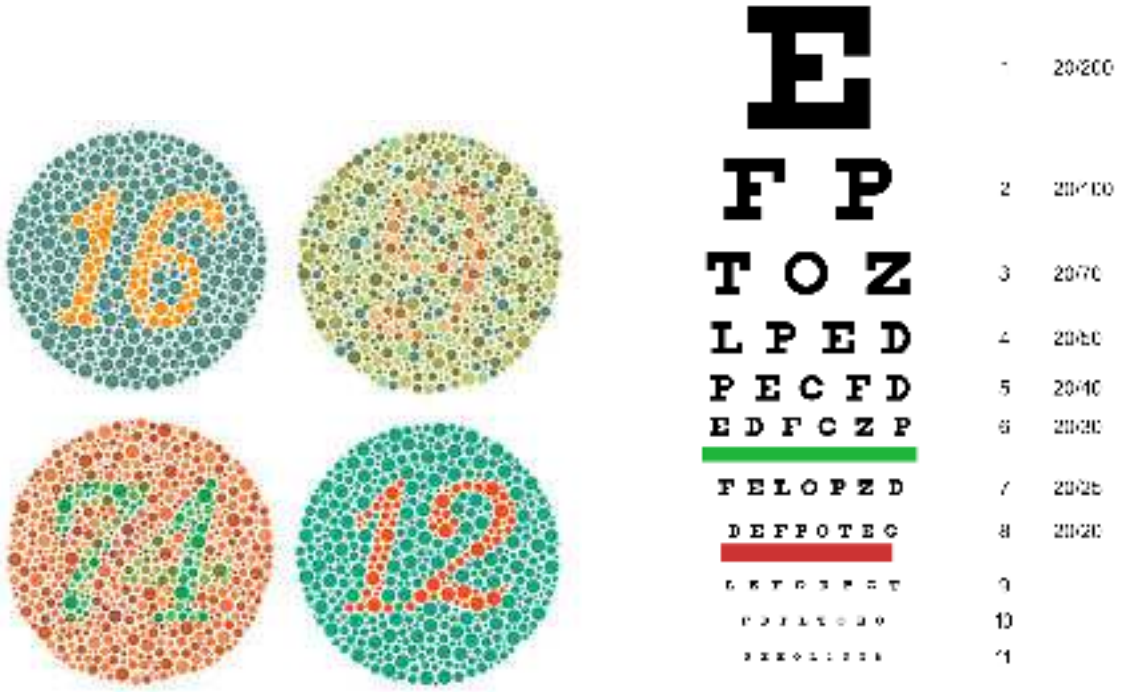


Figure 6.3: Recording of eye movement (saccade, fixation) of the left eye with the Cambridge Technology EyeTracker device during observation of a video sequence.

The number of subjects which have participated in the experience was 21 with age from 20 to 44 years old.

- All subjects have signed a consent form for the use of their anonymous data for research. These forms are safety stored in LaBRI.
- The gaze tracking data are anonymous.
- Subjective test (see figure 6.4) : pure Snellen [119] and Ishihara [50]. We have not identified defective subjects in our volunteers.



(a) Ishihara test (b) Snellen test

Figure 6.4: Snellen [119] and Ishihara [50] tests.

The experimental protocol was the following:

1. The eye tracker was positioned at a distance of 80cm from the screen which size was of 21 inches in diagonal. The eye-tracker was "chin-rest".

2. The instructions to the subjects corresponded to the free viewing conditions without any predefined visual task. Before viewing the videos, the subject was instructed like "Please, watch the video".

3. The examination was carried out in two stages. Initially two series of "normal" or non-degraded videos were presented, then, in a second time two series of "degraded" videos were shown to the subjects.

The visual protocol content for each sequence is illustrated in figure 6.5 and all parameters are summerized in table 6.1.

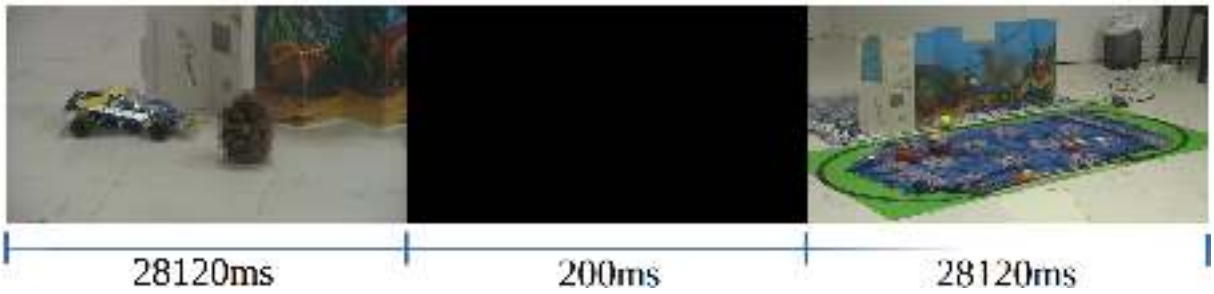


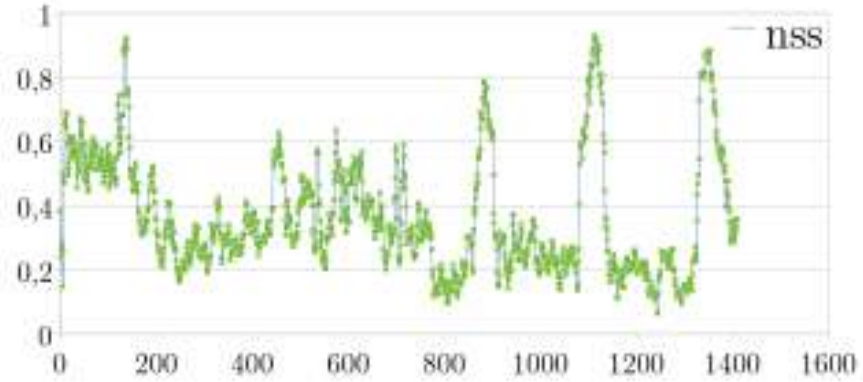
Figure 6.5: Visual protocol content : Sequence of "degraded" videos

Table 6.1: Experiment protocol.

	Features	Values
Video	Video resolution	1280×720
	Video format	2D
	Number of videos	2
Subjective test	Observer distance	80
	Environment	ITU-R BT.500-11
	Duration	56 *second*
	Pre screening	Snellen, Ishihara
Observers	Number of observers	21
	Age : Mean [Range]	26 [20 44]years
	Male / Female repartition	14/8
Eyetracker	Eyetracker	HS-VET
	Eyetracker model	mono ocular
	Eyetracker acquisition frequency	$250Hz$
Display	Display model	HP LP2475w
	Display resolution	1920×1200

Comparaison of saliency maps between normal and degraded sequences

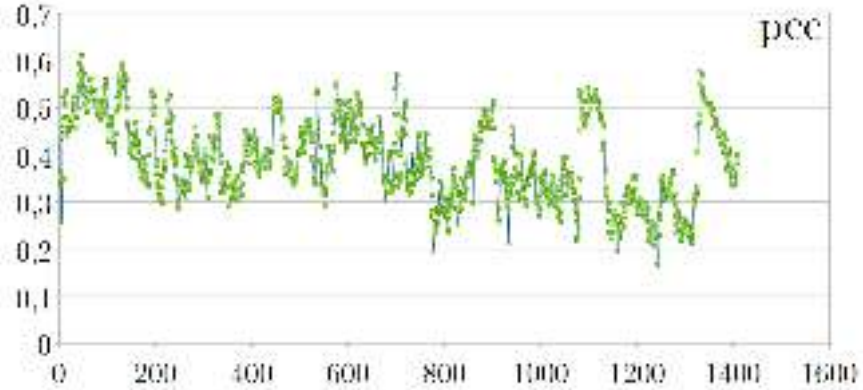
Before conducting experiments with patients, it is important to analyze if the induced degradations attract attention of normal control subjects with regard to non-degraded natural visual content. Hence we will compare subjective saliency maps on a both normal video sequence and corresponding degraded one, frame by frame. Amongst a variety of metrics for comparison of saliency maps those ensuring a simple interpretation of results were chosen. These metrics are: the Pearson correlation coefficient (PCC) and "Normalized scanpath saliency" (NSS) [67] (see 1.3.3).



(a) NSS variations

	NSS_metric
normVSdeg	$0,36522 \pm 0,18139$
normVSdeg-withoutPeak	$0,27335 \pm 0,15923$

(b) mean NSS metric



(c) PCC variations

	PCC_metric
normVSdeg	$0,38843 \pm 0,08566$
normVSdeg-withoutPeak	$0,32412 \pm 0,14740$

(d) mean PCC metric

Figure 6.6: Variations of NSS and PCC metrics during the comparaison of GFDM created for both sequences in the psycho-visual experiment.

Analysing the values of the NSS and PCC metrics when comparing the salient areas of the normal sequence with the degraded sequence, two points were stated (see figure 6.6). First, the values of NSS and PCC are low, that means a large difference between the fixations on areas of the normal sequence and the degraded one by normal control subjects. Therefore, the artificially induced degradations are valuable for further experiments on the patients. Second, we observe four peaks in the sequence (first peak from frame #118 to

#147, second peak from frame #865 to #902, third peak from frame #1081 to #1134, and the fourth one from frame #1327 to #1391). These peaks mean a good correspondence between what attracts attention in the normal sequence and in the degraded one. The frames at these pics correspond to new significant objects entering in the scene. Therefore, normal subjects are attracted by significant objects. They pursue to the new objects dropping the degradation. The mean value of NSS metric without the consideration of the peaks is $0,27335$ (see table (b) of figure 6.6). We can conclude that normal people are attracted by both areas: the significant items and degradation.

6.3 Deep model for study of neuro-degenerative diseases : Mixed model and Merged model

A Deep CNN requires a large amount of data for training. It is impossible to produce such amount of data in the scenario of experiments with patients. Elderly subjects are not able to watch a large amount of visual data in the conditions of eye-tracking experiment. Hence, we try to predict visual attention on degraded sequences with a model trained on a large amount of publicly available data.

When the training samples of source domain are different from the training samples of the target domain, we are in front of the problem of different data distributions. In saliency prediction, a supervised learning approach tries to simulate the sensitivity of HVS to primary features such as contrasts, color saturation and others [125], [10], [104] . One should have expected that if trained on one database (in our case Hollywood) the same model can be successfully applied to another database. Indeed, HVS neurons are sensitive to the same "relative" features. This was our hypothesis in this chapter. Nevertheless, the experiments show that this assumption is not hold. Our explanation is that when observing a content humans interpret it. Here, we try to "fine-tune" a pretrained model on a very specific and small database. We therefore resort to transfer learning. The only public large videos with gaze fixations is HOLLYWOOD2 [88] [89], it is described in chapter 3.6.

As defined in our previous study [27], we modify the Stochastic Gradient Descent (SGD) [12] algorithm used in the learning of Deep CNN parameters. We transfer the model learned on large dataset to the small one. Hence, we start the learning on the small dataset with the best deep CNN parameters already learned on the large database, instead of the random initialization from a gaussian distribution.

With its classical initialization, our transfer method presents the starting of learning from the best weight matrix for each layer of the Deep CNN. In the contrary of the work

of Bengio [3] which uses the transfer of just the three first convolutional layer. The very few available data for training presents the reason to transfer learning of each Deep CNN layers.

Once the binary classification problem : "salient" "Non-salient" has been solved for regions, we need to build predicted saliency maps upon these decisions. (see chapter 3 for more details)

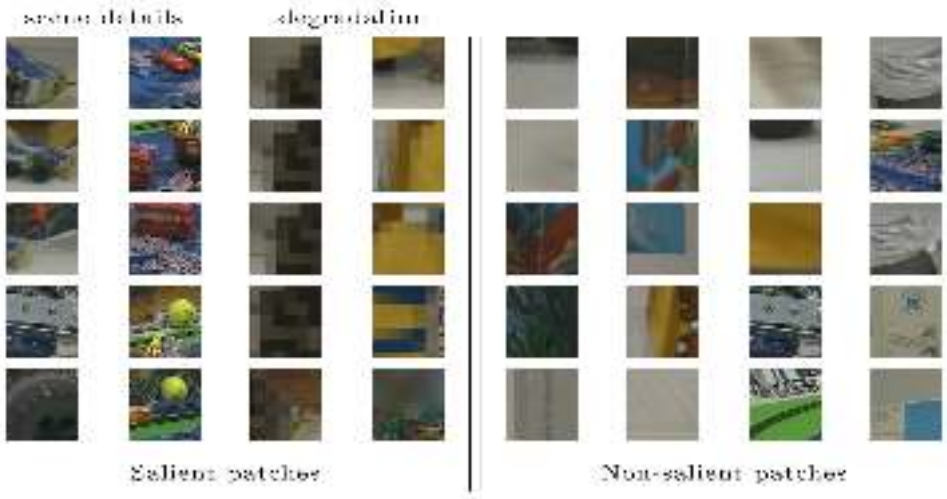
In order to resolve the limitation of the number of frames of produced videos, the data augmentation technique was used to increase the number of salient patches. Since object or area that attracts human gaze is never precisely centered one point in the frame, the translation of the center is required. With the purpose of expanding the variability of the salient class in training dataset, we choose to move the center of salient patch of 5 pixels twice in each direction. The results of this training are presented in the next section together with the benchmarking of proposal saliency prediction model.

6.3.1 Mixed model

As mentioned in LYLO project [123] and proved in our experiment described in section 6.2.2, normal control subjects are attracted by salient areas. The latter can contain an area that contains scene details important for its understanding (a contrast, a moving object ...) or an area that contains a designed degradation. Hence, our first approach consists on mixing the two kinds of salient areas to train one model, we call it "Mixed model".

According to the same approach used to select patches and then to create the dataset for training and validation (see section 3.3.1) , the dataset on degraded sequence for "Mixed model" was created. Here, salient patches can contain: i) degraded area. ii) object of interest. Hence, the two kinds of salient patches were mixed together and the Deep CNN network was trained : "Mixed model". Following table 6.2 presents some examples of patches selected from the degraded sequence to train the "Mixed model".

Table 6.2: Data from degraded sequence to train "Mixed model"



Mixed model is then the combination of all kind of saliency degradations or natural attractors in the training data. It uses exacty the same architecture ChaboNet as described in section 3.4.2 with $4k$ configuration, that is training RGB values and residual motion energy (see section 3.4.1).

6.3.2 Merged model

According to LYLO project [123] a degraded area is salient for normal subject. This is clear in the first experiment in section 6.2.2. Hence, a salient area can be either an area that attract human gaze either a degraded area. The idea here is to create two separated data sets for each kind of salient areas. The first one " NormalInterest" data set is built with reference to GFDM map on normal video sequence without any degradation. The second data set is designed " DegradedInterest". Here, the degraded patches were built upon the mask of degradation. These degraded patches were labled as "saillent". Table 6.3 and 6.4 present an overview of " NormalInterest" and " DegradedInterest" data sets.

Table 6.3: Extract of " NormalInterest" data set.

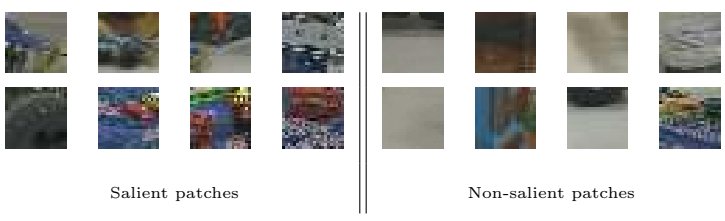
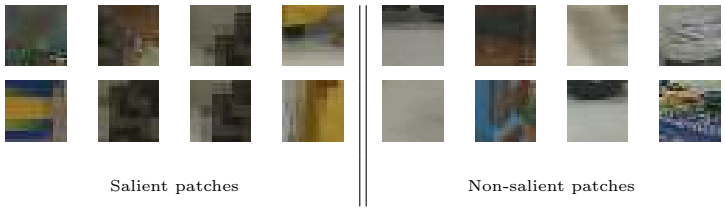


Table 6.4: Extract of " DegradedInterest" data set.



"MergeDinTraning" : Fusion of Normal and Degraded saliency model in training step

Since we have two kinds of interest areas (degradations and normal human gaze attractors), a "ChaboNet" architecture of each kind of interest areas was proposed . The input dataset was 4k configured (RGB values with residual motion energy). The networks are completely identical "seamese" and joint in fully connected layers. The only condition is that two input data images have to be in the same category (salient or Non-salient) . The fusion lies in the last fully connected "FC" layer; a concatenation layer combines these "FC" layers from each single network (see Fig 6.7).

"MergeDinPrediction" : Fusion of Normal and Degraded saliency model at prediction step

The idea in this proposed model, is to make a logical operation on the decision results (Softmax) of the two independent networks, as shown in Fig 6.8. Note that in this approach, firstlly each network was trained separately, then the fusion operation (logical OR) was applied for each forward input data in test step.

6.4 Saliency Generation for Mixed model

Training Deep CNN on degraded sequences from scratch is not thinkable because of the need of a very large database. Therefore, our first idea was to learn the prediction model on a large base (Hollywood2), and use the best model to predict salient areas on the degraded sequence. Sensitivity of HVS to contrasts, color saturation and other low-level features is maintained. Therefore, the predictive power of thus obtained model would be sufficiently good. It was not the case. Despite of the use of a better performing model "Deep saliency RGB8k", with seven kinds of contrasts, residual motion and RGB values, the results of comparison with the ground truth GFDM map were very poor. Indeed, the PCC metric was 0.155 ± 0.069 and NSS was 0.173 ± 0.081. Analyzing predicted

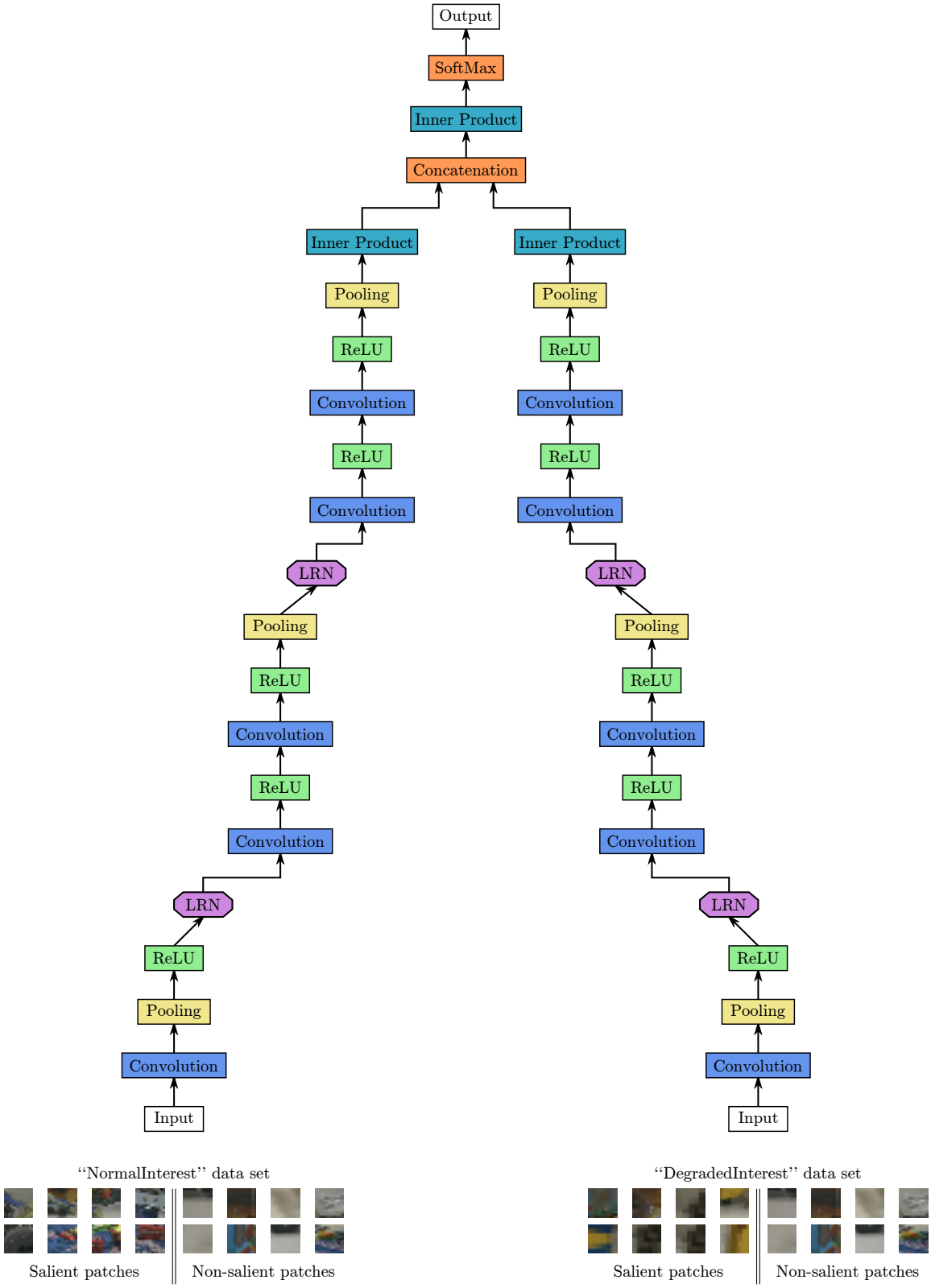


Figure 6.7: Architecture of "MergeDinTraning" model

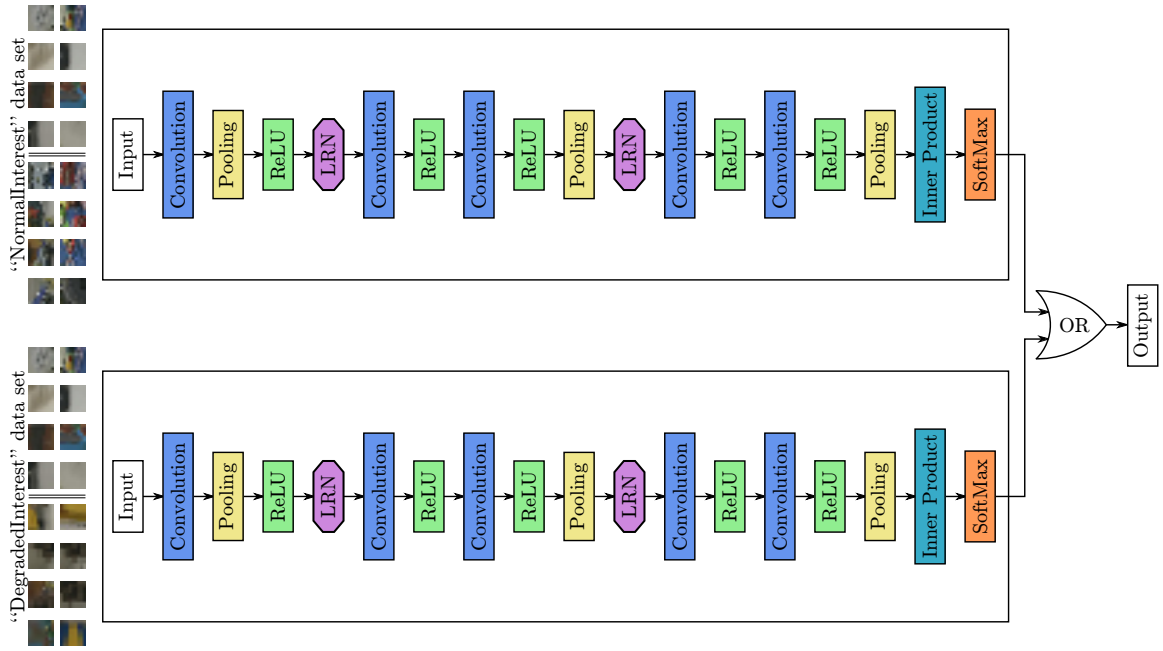


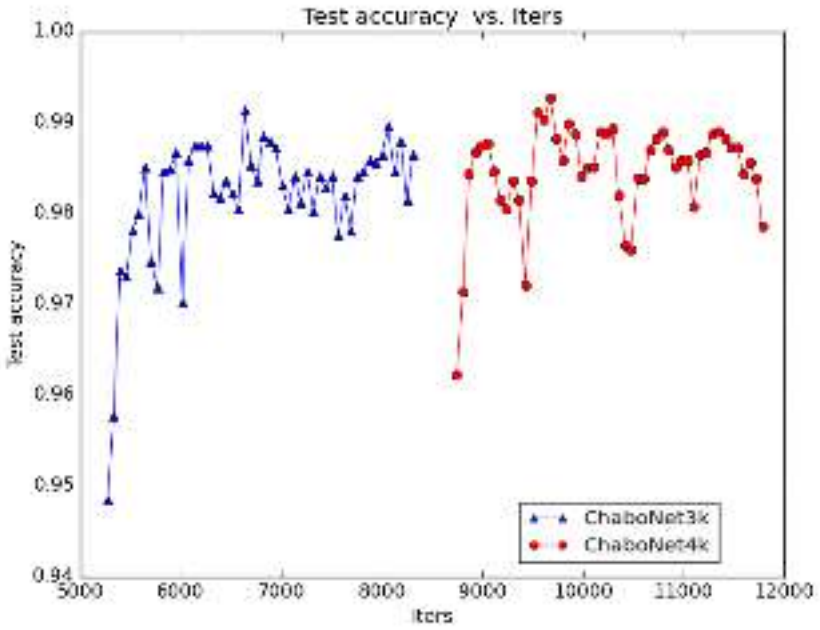
Figure 6.8: Architecture of "MergeDinPrediction" model

maps frame by frame, we discovered that the contrasted degradations were predicted rather satisfactory. Indeed, they are similar to the natural contrasts in video frames. Nevertheless, the intentionally blurred areas were poorly predicted. Such kind of areas was not "seen" by the network in the training data. Thus the adaptation of the models is the must.

6.4.1 Results of transfer learning on Mixed model

Frames available for the learning of the "Mixed model" are very few, just 1404 frames are in our disposal. We have divided them into three sub groups "train", "validation" and "test". From the 939 frames of train, 16028 salient and Non-salient patches were selected. After the creation of the dataset as described in Section 6.3.1 for "Mixed model", the $ChaboNet3k$ and $ChaboNet4k$ models were learned. Transfer learning approach presented in section 5 has been applied. The best $ChaboNet3k$ model trained on the Hollywood dataset is found at the iteration 5214. The learning of the $ChaboNet3k$ for the "Mixed model" was started from the iteration 5214 by transferring the best learned model parameters pretrained on the large dataset. The best model $ChaboNet4k$ learned on HOLLYWOOD2 dataset was found at the iteration 8690, hence, the learning of the $ChaboNet4k$ for the Mixed model on our degraded sequences was started from this iteration (see figure 6.9).

Results summarized in figure 6.9 show the importance of accuracy which attained



(a) Accuracy vs iterations

	$ChaboNet3k$	$ChaboNet4k$
$min_{(\#iter)}$	$94,84\%_{(\#5270)}$	$96,22\%_{(\#8742)}$
$max_{(\#iter)}$	$99,14\%_{(\#6634)}$	$99,27\%_{(\#9672)}$
$avg \pm std$	$98,18\% \pm 0,745$	$98,46\% \pm 0,542$

(c)The accuracy results

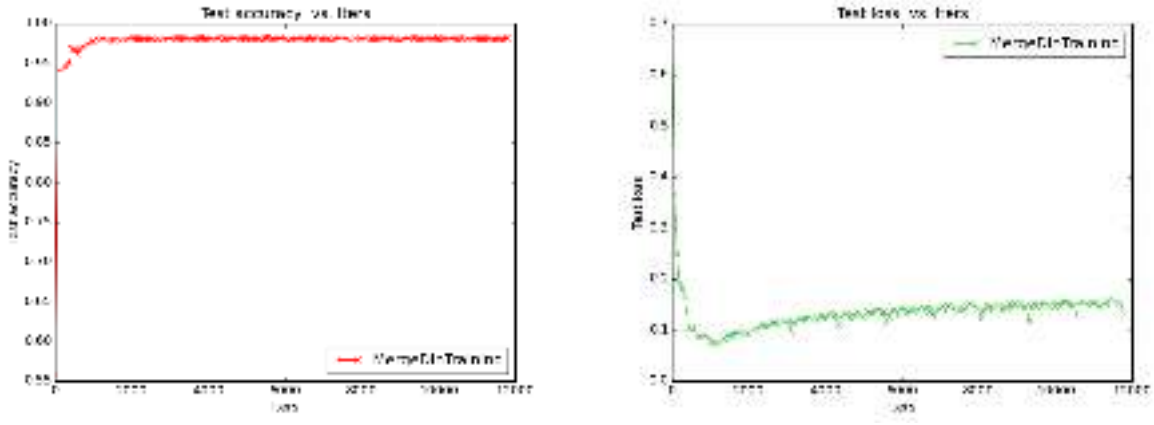
Figure 6.9: Learning of features - Accuracy vs iterations of $ChaboNet3k$ and $ChaboNet4k$ for the "Mixed model".

99.27% at the iteration 9672 for the $ChaboNet4k$ model. We can state that $ChaboNet4k$ outperforms other model in terms of mean accuracy. Nevertheless the gain is not strong.

6.4.2 Results of transfer learning on Merged model

Training of MergeDinTraining

To evaluate the proposed MergeDinTraining architecture, the network was trained to predict the probability of a local region to be salient. The model uses RGB values and the normalized energy of residual motion as input. The MergeDinTraining model was learned from scratch. In the following, obtained results were summarized in figure 6.10.



(a) accuracy curve (b) Loss curve

Figure 6.10: Learning of features - Accuracy and loss vs iterations of the MergeDinTraining model.

The results of learning experiments on NormalInterest 6.3 and DegradedInterest 6.4 data sets yield the following observations:

i) The results of accuracy are rather good : average accuracy is about 97.74% (see table 5.7).

ii) The accuracy curve (figure 6.10 (a)) and the corresponding loss curve (figure 6.10(b)) show that the best trained model reached 98.33% of accuracy with the smallest loss (at the iteration #3100 see table 6.5). Thus, it does not present an over-fitting situation.

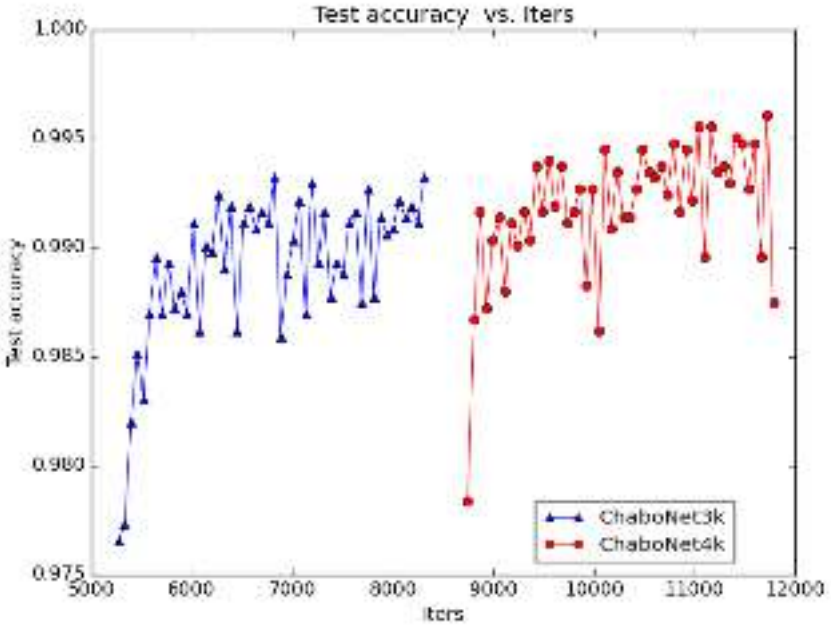
Table 6.5: The accuracy results on learned MergeDinTraining model.

$min\_Accuracy_{(\#iter)}$	$55.80\%_{(\#0)}$
$max\_Accuracy_{(\#iter)}$	$98.33\%_{(\#3100)}$
$avg\_Accuracy \pm std$	$\%97.74 \pm 3.115$

Training of MergeDinPrediction

Figure 6.11 illustrates the variations of the accuracy along iterations of ChaboNet3k and ChaboNet4k networks for DegradedInterest 6.4 data set. To overcome the lack of data, the learning was transferred from the best obtained models on "HOLLYWOOD" data set. The gain of using 4k against 3k as input to the deep CNNs is about 0.2% in terms

of mean accuracy. The best model is obtained at the iteration #11718 with an accuracy of 99.60%.



(a) Accuracy vs iterations

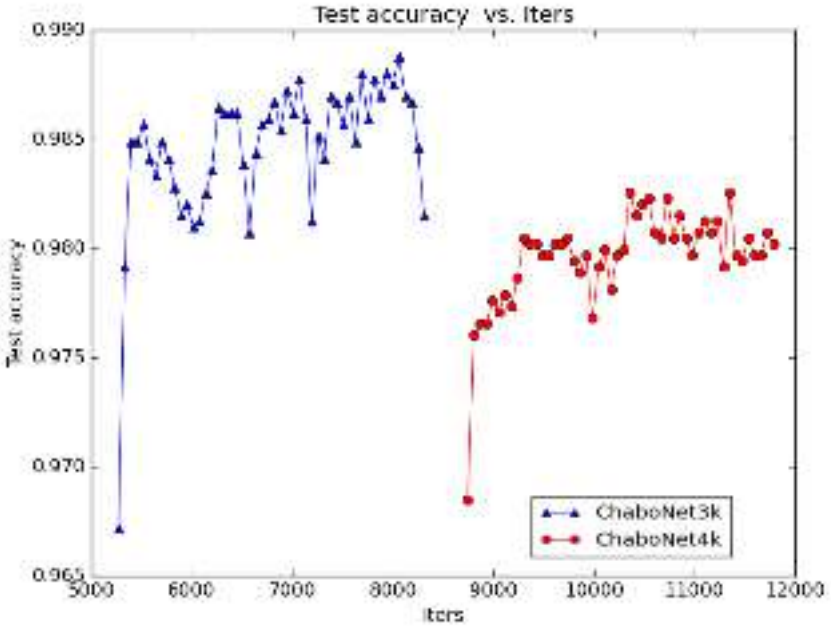
	$ChaboNet3k$	$ChaboNet4k$
$min_{(\#iter)}$	$97.65\%_{(\#5270)}$	$97.83\%_{(\#8742)}$
$max_{(\#iter)}$	$99.32\%_{(\#6820)}$	$99.60\%_{(\#11718)}$
$avg \pm std$	$98.9\% \pm 0.355$	$99.18\% \pm 0.3048$

(c)The accuracy results on degraded model

Figure 6.11: Learning of features - Accuracy vs iterations of $ChaboNet3k$ and $ChaboNet4k$ for the "DegradedInterest" data set.

Figure 6.12 illustrates the variations of the accuracy along iterations of ChaboNet3k and ChaboNet4k networks for NormalInterest 6.3 data set. To overcome the lack of data, the learning was transferred from the best obtained models on "HOLLYWOOD" data set. From the plots (a) in figure 6.12, we can see that the $ChaboNet4k$ model is little less efficient than $ChaboNet3k$ model. It is not surprising, the salient patches are predicted by our method according to each visual task: on the Hollywood data set the subjects are instructed to observe actions. They are attracted by the dynamic content of the visual scene. Hence residual motion is important in the global model. In NormalInterest data

set, the subjects are interested in specific objects be they moving or not. Hence, the spatial appearance is important.



(a) Accuracy vs iterations

	$ChaboNet3k$	$ChaboNet4k$
$min_{(\#iter)}$	$96.71\%_{(\#5270)}$	$96.84\%_{(\#8742)}$
$max_{(\#iter)}$	$98.88\%_{(\#8060)}$	$98.25\%_{(\#10354)}$
$avg \pm std$	$98.46\% \pm 0.333$	$97.95\% \pm 0.222$

(c)The accuracy results

Figure 6.12: Learning of features - Accuracy vs iterations of $ChaboNet3k$ and $ChaboNet4k$ for the "NormalInterest" data set.

6.5 Comparaison of predicted saliency maps on degraded sequence

In literature, different evaluation metric were used to determine the rate of similarity between the saliency maps and the gaze fixations of subjects. Three criteria allowing an easy interpretation of results were chosen. These criteria are: the correlation coefficients 'CC' used in various areas to assess the similarity of two distributions, the receiver effi-

ciency (AUC) for evaluating the quality of a prediction, and NSS "Normalized scanpath saliency" which is defined to compare the salient areas determined by a model with areas observed by the subjects [85].

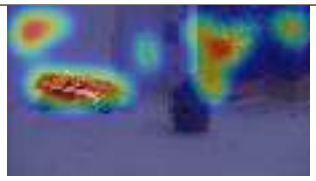
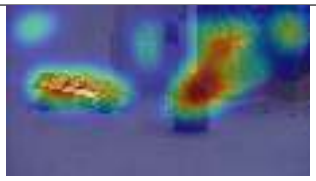
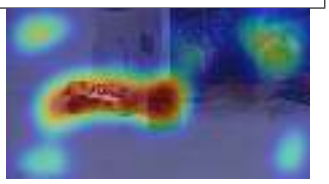
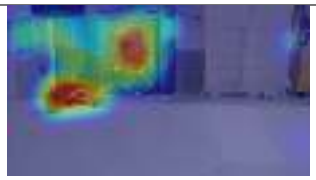
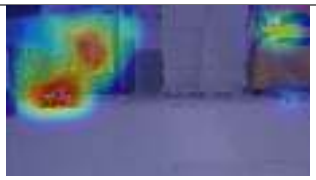
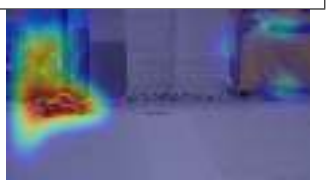
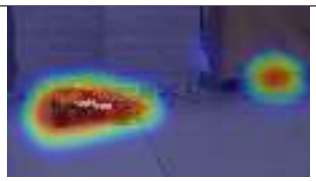
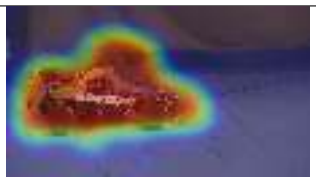
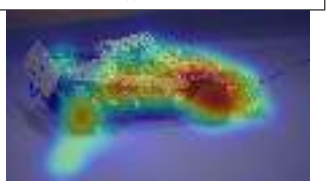
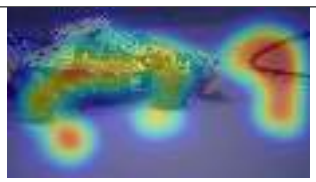
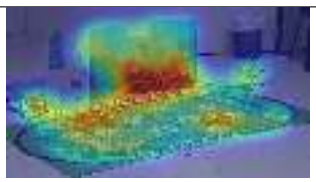
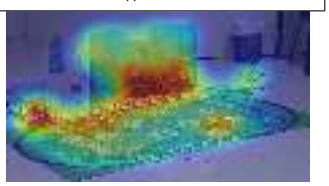
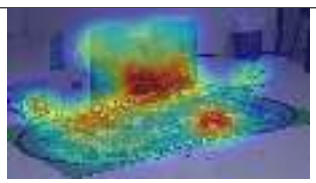
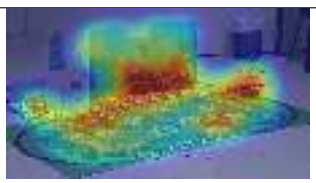
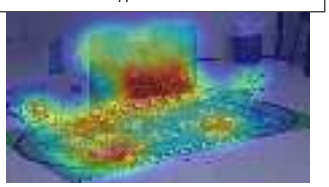
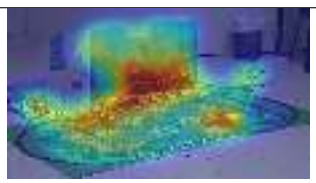
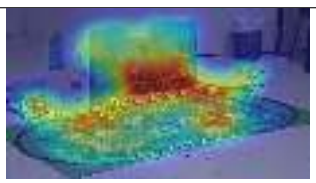
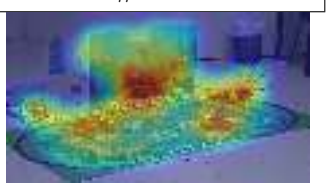
The average and the standard deviation of the NSS, AUC and CC metric were computed for all test frames (see table 6.6). The proposed model was compared with the static model of Itti [52] and the dynamic model of Seo [113] which are the common benchmarks in literature. For the AUC metric, the proposed "Mixed model" outperforms the Seo model. Nevertheless, the latter have a quite better mean value of NSS and CC but the standard deviation is strong that reflects that our proposed model presents more stable results. The proposed "Mixed model" outperforms in mean value of NSS and CC metric the Itti model. Nevertheless, the latter outperforms our "Mixed model" in mean AUC metric. The proposed "Mixed model" outperforms other proposed models MergeDinTraining and MergeDinTraining.

Table 6.6: comparison with AUC, NSS and CC metric of gaze fixations 'GFM' vs predicted saliency of Mixed model, Itti model and Seo model for the 235 test frame of degraded sequence

Metric	GFM vs Mixed model	GFM vs MergeDinTraining	GFM vs MergeDinPrediction	GFM vs Itti model	GFM vs Seo model
AUC	0.756 ± 0.227	$0,58 \pm 0,218$	0.494 ± 0.163	0.769 ± 0.163	0.630 ± 0.216
NSS	1.029 ± 0.990	$0,43 \pm 1,292$	-0.0529 ± 0.8086	0.952 ± 0.780	1.185 ± 2.65
CC	0.042 ± 0.041	$0,02 \pm 0,058$	-0.0022 ± 0.0344	0.040 ± 0.032	0.046 ± 0.096

Next table 6.7 presents some examples of predicted saliency map with the proposed "Mixed model". For these frames, a very interesting value of AUC compared to gaze fixation was obtained. We can see that our model predict the objects of interest specified by contrast or residual motion likewise the intentionally degraded area.

Table 6.7: Examples of predicted saliency map with *ChaboNet4k* of proposed "Mixed model"

FRAME	#38	#68	#146
			
AUC	0.957	0.935	0.99
FRAME	#230	#260	#326
			
AUC	0.974	0.988	0.828
FRAME	#416	#506	#584
			
AUC	0.870	0.912	0.991
FRAME	#608	#734	#830
			
AUC	0.971	0.999	0.961
FRAME	#878	#920	#1016
			
AUC	0.887	0.871	0.907
FRAME	#1094	#1142	#1388
			
AUC	0.924	0.978	0.902

6.6 Conclusion

Hence, in this research we produced and we validated a video content for psycho-visual experiments with dementia patients. Furthermore, we have built a reference model for normal subjects observing intentionally degraded content. The model was built on the basis of Deep CNNs. In a medical study, we face a typical situation of a very small database. Even very good models applied on the new unseen content in the same saliency prediction task not performing well, we proposed a transfer learning scheme. Here we fine-tuned the initial model. Several models were proposed in this chapter.

Mixed model, which is trained on a very low amount of data (our degraded sequences) was developed. It gives very good results, such as a predictive accuracy of 99.27% due to the efficiency of transfer learning method. From what we can see in the obtained results, many frames achieved 0.9 of AUC value. This means that our model predicts well both gaze attraction by semantic objects and unusual degradations.

Two merged models were learned: MergeDinTraining siamese network model and MergeDinPrediction which implement the fusion of two separately learned models by logical OR operation (max).

General conclusion and perspectives

In this work, we were interested in prediction of visual saliency in video content with the new classification tools such as Deep Convolutional Neural Networks. The target application of this research was building of a model for prediction of saliency of regions in video for studies of attention of patients with neuro-degenerative diseases. To build an efficient model, we explored different aspects of these supervised classifiers in the problem of saliency prediction such as
- design of an adequate architecture of a Deep CNN;
- studies of possible input layers of the architecture on the basis of domain knowledge, such as sensitivity of human visual system to contrasts, colour and residual motion in dynamic content;
- sensitivity of these classifiers to the noise in training data;
- efficient initialization of parameters by transfer learning;
- fusion of classification results in the problem of recognition of various kinds of intentional degradations designed for studies of attention of patients.

To explore a video, we focus our attention on certain salient regions whose the movement and the semantic aspect of the object-of interest represent visual attractors. For this purpose in collaboration with medical researchers we have designed specific degradations in video succeptible to attract attention, designed and conducted psycho-visual experiment and studied the reaction of normal control subjects on these degaradations for groundtrouthing of prediction model. This small database was used in the present work together with larger video databases with available gaze fixation recordings such as Hollywood large-scale data base or well-known in video quality assessment community IRCCyN database. We did mastering of these databases by content selection for training and validation of our models. Benchmarking of our contributions with regard to the available ground trouth, but also with regard to classical visual saliency prediction models form the litterature was performed. In the following paragraphs we will focus on our contributions and propose perspectives of the present research.

Contributions of this PhD and their Assessment

Firstly, a model with four channels based on the colors R, G, B and motion was proposed. Then this model was enriched with seven other channels summarizing the different kinds of contrast already studied for the saliency prediction. Throughout this thesis, we have used databases that collect information on the human gaze recorded through a psychovisual experiments. The use of this information allowed us to define the target class of salient regions.

- ChaboNet: a Deep CNN architecture built on the basis of AlexNet [63]. The challenge here was to design an architecture which would not be "too deep" in order to have reasonable times of training and also to limit numbers of learning parameters in order to get a stability. Our contribution was in adding supplementary architectural patterns of convolution and non-linearity layers before pooling layers with the goal to increase the "expressivity" of features. The latter is important in saliency prediction as HVS is sensitive to contrasts and singularities both spatial and temporal. We have also reduced the number of filters to learn. The benchmarking of the proposed architecture with regard to base-line AlexNet architecture has shown a slight increase of accuracy in prediction of saliency of regions in video.

- A specific input data layer of Deep CNN for visual saliency prediction. First, the use of eye fixation dense map in training deep CNNs models ensures the combination of both bottum-up and top-down saliency cues. Second, for video processing, the temporal cues are mainly prevalent to detect salient region. The experiments have shown promising results. Furthermore, to explore the domain knwoled on the sensitivity of HVS to specific contrasts we have conducted experiments using seven kinds of contrasts as input of the deep CNN. This allowed us to be sure of certain choices of the model and to limit the input model on temporal component with the RGB values.

- Sensitivity of deep CNNs to noise in training data. We have stated the noise in the automatic production of training data in video with reference Gaze Fixation Density Maps only. And we have proposed training data selection process on the basis of visual content production rules reducing the noise. Despite a systematic study of the influence of noise in the input data was out-of-scope of our research, we have shown that filtering noise in training data allows for increasing of accuracy of prediction.

- Transfer learning with deep CNNs. A typical situation in real-life applications of

Deep learning, specifically in medical research domain, is the limited number of training data. Hence, we have proposed and tested a method of transfer learning, as a fine-tunning of parameters initialized with training on a large dataset in the same saliency prediction problem. This method was fully studied and experimented on three small datasets. Finally, we have applied it to a task with very small amount of training data in the problem of prediction of reference normal control visual attention for studies of neuro-degenerative diseases.

– Generation of saliency map: we proposed a specific method to generate the final saliency map. Inspired from GFDM with Wooding's method, we used the probability responses of deep CNNs model to create a saliency map with the same size of input frames. The codes were optimized with a parallel algorithm that reduces the time of generation of saliency maps.

– eye-tracking experiences : an eye-tracking experiment was designed for testing patients with neurodegenerative diseases. First of all specific video content with intended degradation was produced in collaboration with researchers in medicine. Then the experiment was conducted on healthy volunteers in free viewing conditions.

We have been able to draw several conclusions such as normal subjects are attracted by significant objects. They pursue to the new objects dropping the degradation. Experimenting with our prediction model, we have designed fusion strategy for learning and prediction of different kinds of degradations.

Last but not least for a PhD in Computer Science, a total of ten softwares and scripts were developed for this research project using different opensource frameworks or matlab.

In this work we have not systematically quantized the performances of our approach in terms of execution time, for the reason of heterogeneous equipements we have used along this research. Nevertheless, a systematic tracking of accuracies along the iteration of training of our models has allowed us to drastically (order of 10) reduce the number of iterations compared to the state-of-the-art research.

Perspectives

This work opens many perspectives which can be envisaged either as its improvement or its direct extension or as requiring extensive and longer-term studies. Deeper exploitation

of the model possibilities can be made by boosting the *ChaboNet4k* with a step of fine-tuning from other trained models in particular by net surgely operation.

Using Fully convolutional networks for saliency prediction on natural videos, presents a new research perspective which we would like to explore.

Furthermore, temporal consistency of saliency maps can also be improved using other kinds of architectures, than CNNs.

In conclusion, we believe that the proposed saliency model using deep CNNs has a very good application perspective, especially in neurodegenerative diseases diagnostics and several other saliency prediction applications such as video compression, watermarking and selective indexing of visual content.

Bibliography

[1] M. Abadi, A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro, G. S. Corrado, A. Davis, J. Dean, M. Devin, S. Ghemawat, I. Goodfellow, A. Harp, G. Irving, M. Isard, Y. Jia, R. Jozefowicz, L. Kaiser, M. Kudlur, J. Levenberg, D. Mané, R. Monga, S. Moore, D. Murray, C. Olah, M. Schuster, J. Shlens, B. Steiner, I. Sutskever, K. Talwar, P. Tucker, V. Vanhoucke, V. Vasudevan, F. Viégas, O. Vinyals, P. Warden, M. Wattenberg, M. Wicke, Y. Yu, and X. Zheng. TensorFlow: Large-scale machine learning on heterogeneous systems, 2015. Software available from tensorflow.org.

[2] J. Ahmad, Sajjad M., Mehmood I., and Baik S. W. SiNC: Saliency-injected neural codes for representation and efficient retrieval of medical radiographs. *PLOS ONE*, 12(8):1–32, 2017.

[3] M. Ammar, M. Mitrea, M. Hasnaoui, and P. Le Callet. Visual saliency in MPEG-4 AVC video stream. volume 9394, pages 93940X–93940X–11, 2015.

[4] N. K. Archibald, S. B. Hutton, M. P. Clarke, U. P. Mosimann, and D. J. Burn. Visual exploration in Parkinson's disease and Parkinson's disease dementia. *Brain journal de neurologie*, 2013.

[5] M. Assens, K. McGuinness, X. Giró, and N. E. O'Connor. SaltiNet: Scan-path Prediction on 360 Degree Images using Saliency Volumes. *CoRR*, abs/1707.03123, 2017.

[6] Y. Bengio. Deep Learning of Representations for Unsupervised and Transfer Learning. In *Proceedings of the 2011 International Conference on Unsupervised and Transfer Learning Workshop - Volume 27*, UTLW'11, pages 17–37. JMLR.org, 2011.

[7] Y. Bengio, P. Lamblin, D. Popovici, and H. Larochelle. Greedy Layer-Wise Training of Deep Networks. In P. B. Schölkopf, J. C. Platt, and T. Hoffman, editors, *Advances in Neural Information Processing Systems 19*, pages 153–160. MIT Press, 2007.

[8] J. Benois-Pineau, S. Anthoine, C. Morand, J. P. Domenger, E. Debreuve, W. Bel Haj Ali, and P. Piro. Scalable Indexing of HD Video. In Mislav; Kunt Murat Mrak, Marta; Grgic, editor, *High-Quality Visual Experience*, Signals and Communication Technology, pages 497–524. Springer, June 2010.

[9] J. Bergstra, O. Breuleux, F. Bastien, P. Lamblin, R. Pascanu, G. Desjardins, J. Turian, D. Warde-Farley, and Y. Bengio. Theano: a CPU and GPU math expression compiler. In *Proceedings of the Python for Scientific Computing Conference (SciPy)*, 2010. Oral Presentation.

[10] A. Borji and L. Itti. State-of-the-art in Visual Attention Modeling. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(1):185–207, Jan 2013.

[11] L. Bottou. On-line Learning in Neural Networks. chapter On-line Learning and Stochastic Approximations, pages 9–42. Cambridge University Press, New York, NY, USA, 1998.

[12] L. Bottou. Stochastic Gradient Tricks. In *Neural Networks, Tricks of the Trade, Reloaded*, pages 430–445. Springer, 2012.

[13] L. Bottou and O. Bousquet. The Tradeoffs of Large Scale Learning. In *Proceedings of the 20th International Conference on Neural Information Processing Systems*, NIPS'07, pages 161–168, USA, 2007. Curran Associates Inc.

[14] H. Boujut, J. Benois-Pineau, T. Ahmed, O. Hadar, and P. Bonnet. No-reference video quality assessment of H.264 video streams based on semantic saliency maps. volume 8293, pages 82930T–82930T–9, 2012.

[15] H. Boujut, R. Mégret, and J. Benois-Pineau. Fusion of Multiple Visual Cues for Visual Saliency Extraction from Wearable Camera Settings with Strong Motion. In *Computer Vision - ECCV 2012. Workshops and Demonstrations - Florence, Italy, October 7-13, 2012, Proceedings, Part III*, pages 436–445, 2012.

[16] F. Boulos, W. Chen, B. Parrein, and P. Le Callet. Region-of-Interest Intra Prediction for H.264/AVC Error Resilience. In *IEEE International Conference on Image Processing*, pages 3109–3112, Cairo, Egypt, November 2009.

[17] O. Brouard, V. Ricordel, and D. Barba. Cartes de saillance spatio-temporelle basées contrastes de couleur et mouvement relatif. In *Compression et représentation des signaux audiovisuels, CORESA 2009*, page 6 pages, Toulouse, France, 2009.

[18] N. Bruce and J. Tsotsos. Saliency based on information maximization. *In Advances in Neural Information Processing Systems*, (18):155–162, 2006.

[19] V. Buso. *Reconnaissance perceptuelle des objets d'Intérêt : application à l'interprétation des activités instrumentales de la vie quotidienne pour les études de démence*. PhD thesis, 2015. Thèse de doctorat dirigée par Benois Pineau, Jenny Informatique Bordeaux 2015.

[20] F. Chan, I. T. Armstrong, G. Pari, R. J. Riopelle, and D. P. Munoz. Deficits in saccadic eye-movement control in Parkinson's disease. *Neuropsychologia*, 43(5):784–796, 2005.

[21] R. Collobert, S. Bengio, and J. Marithoz. Torch: A modular machine learning software library, 2002.

[22] S. Cubizolle, N. Damon-Perriére, S. Dupouy, A. Foubert-Samier, and F. Tison. Parkinson's disease, l-dopa and express saccades: Superior colliculus dyskinesias? *Clinical Neurophysiology*, 125(3):647–648, oct 2014.

[23] Scott J. Daly. Engineering observations from spatiovelocity and spatiotemporal visual models, 1998.

[24] S. F. Dodge and L. J. Karam. Visual saliency prediction using a mixture of deep neural networks. *CoRR*, abs/1702.00372, 2017.

[25] W. Elloumi, K. Guissous, A. Chetouani, and S. Treuillet. Improving a vision indoor localization system by a saliency-guided detection. pages 149–152, 2014.

[26] U. Engelke, H. Lieu, J. Wang, P. Le callet, I. Heynderickx, H. j. Zepernick, and A. Maeder. Comparative study of fixation density maps. *IEEE Trans. Image Processing*, 22(3):1121–1133, March 2013.

[27] A. Fathi, X. Ren, and J. M. Rehg. Learning to recognize objects in egocentric activities. In *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, pages 3281–3288, June 2011.

[28] M. A. Fischler and R. C. Bolles. Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. *Magazine Communications of the ACM*, 24(6):381–395, 1981.

[29] T. Foulsham and G. Underwood. What can saliency models predict about eye movements? Spatial and sequential aspects of fixations during encoding and recognition. *Journal of Vision*, 8(2):6, 2008.

[30] D. Gao, V. Mahadevan, and N. Vasconcelos. On the plausibility of the discriminant center-surround hypothesis for visual saliency. *Journal of Vision*, (8(7):13):1–18, 2008.

[31] D. Gao and N. Vasconcelos. Integrated learning of saliency, complex features, and object detectors from cluttered scenes. *IEEE Conference on Computer Vision and Pattern Recognition.*, (17), 2005.

[32] C. Garcia and M. Delakis. Convolutional face finder: a neural architecture for fast and robust face detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 26(11):1408–1423, Nov 2004.

[33] M. Geng, Y. Wang, T. Xiang, and Y. Tian. Deep Transfer Learning for Person Re-identification. *CoRR*, abs/1611.05244, 2016.

[34] R. B. Girshick, J. Donahue, T. Darrell, and J. Malik. Region-Based Convolutional Networks for Accurate Object Detection and Segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.*, 38(1):142–158, 2016.

[35] G. Gitchel, P. Wetzel, and M. Baron. Pervasive Ocular Tremor in Patients With Parkinson Disease. *Arch Neurol*, April 2012.

[36] B. E. Goldstein. *Sensation and Perception 8th Edition*. Cengage Learning, 2009.

[37] I. González-Díaz, J. Benois-Pineau, V. Buso, and H. Boujut. Fusion of Multiple Visual Cues for Object Recognition in Videos. pages 79–107, 2014.

[38] I.J. Goodfellow, D. Warde-farley, M. Mirza, A. Courville, and Y. Bengio. Maxout networks. In *In ICML*, 2013.

[39] C. Guo and L. Zhang. A Novel Multiresolution Spatiotemporal Saliency Detection Model and Its Applications in Image and Video Compression. *IEEE Trans. Image Processing*, 19(1):185–198, 2010.

[40] J. Han, L. Sun, X. Hu, J. Han, and L. Shao. Spatial and temporal visual attention prediction in videos using eye movement data. *Neurocomputing*, 145:140–153, 2014.

[41] J. Harel, Ch. Koch, and P. Perona. Graph-based visual saliency. In *Advances in neural information processing systems 19*, pages 545–552. MIT Press, 2007.

[42] K. He, X Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. *CoRR*, abs/1512.03385, 2015.

[43] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. *CoRR*, abs/1512.03385, 2015.

[44] G. E. Hinton, S. Osindero, and Y. W. Teh. A Fast Learning Algorithm for Deep Belief Nets. *Neural Comput.*, 18(7):1527–1554, July 2006.

[45] G. E. Hinton and R. R. Salakhutdinov. Reducing the dimensionality of data with neural networks. *Science*, 313(5786):504–507, July 2006.

[46] B. K.P. Horn and B. G. Schunck. Determining optical flow. *Artificial Intelligence*, 17(1):185–203, 1981.

[47] X. Hou, J. Harel, and Ch. Koch. Image Signature: Highlighting Sparse Salient Regions. *IEEE Trans. Pattern Anal. Mach. Intell.*, 34(1):194–201, 2012.

[48] X. Hou and L. Zhang. Dynamic visual attention: Searching for coding length increments. *In Advances in Neural Information Processing Systems,*, (21):681–688, 2008.

[49] A. D. Hwang, H.C. Wang, and M. Pomplun. Semantic guidance of eye movements in real-world scenes. *Vision Research*, 51(10):1192–1205, 2011.

[50] S. Ishihara. *Tests for colour blindness*. [Tokio], 15th ed. edition, 1917.

[51] L. Itti. CRCNS Data Sharing: Eye movements during free-viewing of natural videos. In *Collaborative Research in Computational Neuroscience Annual Meeting, Los Angeles, California*, Jun 2008.

[52] L. Itti, C. Koch, and E. Niebur. A Model of Saliency-Based Visual Attention for Rapid Scene Analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(11):1254–1259, Nov 1998.

[53] ITU-R. Recommendation 500-11: Methodology for the subjective assessment of the quality of television pictures. ITU-R Rec. BT.500-11, 2002.

[54] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell. Caffe: Convolutional Architecture for Fast Feature Embedding. In *Proceedings of the ACM International Conference on Multimedia, MM '14, Orlando, FL, USA, November 03 - 07, 2014*, pages 675–678, 2014.

[55] A. Joulin, L. van der Maaten, A. Jabri, and N. Vasilache. *Learning Visual Features from Large Weakly Supervised Data*, pages 67–84. Springer International Publishing, Cham, 2016.

[56] S. Karaman, J. Benois-Pineau, V. Dovgalecs, R. Mégret, J. Pinquier, R. André-Obrecht, Y. Gaëstel, and J. F. Dartigues. Hierarchical Hidden Markov Model in detecting activities of daily living in wearable videos for studies of dementia. *Multimedia Tools Appl.*, 69(3):743–771, 2014.

[57] A. Karpathy. Stanford University CS231n: Convolutional Neural Networks for Visual Recognition.

[58] K. C. Kiwiel. Convergence and efficiency of subgradient methods for quasiconvex minimization. *Mathematical Programming*, 90(1):1–25, Mar 2001.

[59] K. Krafka, A. Khosla, P. Kellnhofer, H. Kannan, S. Bhandarkar, W. Matusik, and A. Torralba. Eye tracking for everyone. *IEEE Conference on Computer Vision and Pattern Recognition*.

[60] J. Krause, B. Sapp, A. Howard, H. Zhou, A. Toshev, T. Duerig, J. Philbin, and L. Fei-Fei. *The Unreasonable Effectiveness of Noisy Data for Fine-Grained Recognition*, pages 301–320. Springer International Publishing, Cham, 2016.

[61] A. Krizhevsky. *Learning multiple layers of features from tiny images*. PhD thesis, University of Toronto, 2009.

[62] A. Krizhevsky, I. Sutskever, and G. E. Hinton. ImageNet Classification with Deep Convolutional Neural Networks. pages 1097–1105, 2012.

[63] A. Krizhevsky, I. Sutskever, and G. E. Hinton. ImageNet Classification with Deep Convolutional Neural Networks. In F. Pereira, C.J.C. Burges, L. Bottou, and K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems 25*, pages 1097–1105. Curran Associates, Inc., 2012.

[64] S. S. S. Kruthiventi, V. Gudisa, J. H. Dholakiya, and R. V. Babu. Saliency Unified: A Deep Architecture for simultaneous Eye Fixation Prediction and Salient Object Segmentation. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 00:5781–5790, 2016.

[65] M. Kümmerer, L. Theis, and M. Bethge. Deep Gaze I: Boosting Saliency Prediction with Feature Maps Trained on ImageNet. *CoRR*, abs/1411.1045, 2014.

[66] B. Lai and X. Gong. Saliency guided end-to-end learning for weakly supervised object detection. *CoRR*, abs/1706.06768, 2017.

[67] O. Le Meur and T. Baccino. Methods for comparing scanpaths and saliency maps: strengths and weaknesses. *Behavior Research Methods*, 45(1):251–266, Mar 2013.

[68] O. Le Meur and Z. Liu. Saccadic model of eye movements for free-viewing condition. *Vision research*, 116:152–164, 2015.

[69] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-Based Learning Applied to Document Recognition. Number 86(11), pages 2278–2324, November 1998.

[70] G. Lee, Y. W. Tai, and J. Kim. Deep Saliency with Encoded Low level Distance Map and High Level Features. *IEEE Conference on Computer Vision and Pattern Recognition*, abs/1604.05495, 2016.

[71] G. Lee, Y. W. Tai, and J. Kim. Eld-net: An efficient deep learning architecture for accurate saliency detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PP(99):1–1, 2017.

[72] H. Lee, R. Grosse, R. Ranganath, and A. Y. Ng. Convolutional Deep Belief Networks for Scalable Unsupervised Learning of Hierarchical Representations. In *Proceedings of the 26th Annual International Conference on Machine Learning*, ICML '09, pages 609–616, New York, NY, USA, 2009. ACM.

[73] H. Lee, P. Pham, Y. Largman, and A. Y. Ng. Unsupervised feature learning for audio classification using convolutional deep belief networks. In Y. Bengio, D. Schuurmans, J. D. Lafferty, C. K. I. Williams, and A. Culotta, editors, *Advances in Neural Information Processing Systems 22*, pages 1096–1104. Curran Associates, Inc., 2009.

[74] O. LeMeur, P. LeCallet, D. Barba, and D. Thoreau. A coherent computational approach to model bottom-up visual attention. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28(5):802–817, May 2006.

[75] G. Li and Y. Yu. Visual saliency based on multiscale deep features. *IEEE Conference on Computer Vision and Pattern Recognition*, pages 5455–5463, 2015.

[76] G. Li and Y. Yu. Deep contrast learning for salient object detection. *IEEE Conference on Computer Vision and Pattern Recognition*, 1603.01976, 2016.

[77] G. Li and Y. Yu. Visual Saliency Detection Based on Multiscale Deep CNN Features. *IEEE Transactions on Image Processing,*, 25(11):5012–5024, 2016.

[78] C. Liu. *Beyond Pixels: Exploring New Representations and Applications for Motion Analysis*. PhD thesis, Cambridge, MA, USA, 2009. AAI0822221.

[79] H. Liu, D. Xu, Q. Huang, W. Li, M. Xu, and S. Lin. Semantically-Based Human Scanpath Estimation with HMMs. 12 2013.

[80] N. Liu, D. Han, J.and Zhang, S. Wen, and T. Liu. Predicting eye fixations using convolutional neural networks. *IEEE Conference on Computer Vision and Pattern Recognition*, pages 362–370, 2015.

[81] Z. Liu, P. Luo, S. Qiu, X. Wang, and X. Tang. DeepFashion: Powering Robust Clothes Recognition and Retrieval With Rich Annotations. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

[82] Z. Liu, X. Zhang, S. Luo, and O. Le Meur. Superpixel-Based Spatiotemporal Saliency Detection. *IEEE Transactions on Circuits and Systems for Video Technology*, 24(9):1522–1540, 2014.

[83] Y. Lu, Z. Li, X. Zhang, B. Ming, J. Jia, R. Wang, and D. Ma. Retinal nerve fiber layer structure abnormalities in early Alzheimer's disease: evidence in optical coherence tomography. *Neurosci Lett*, (480):69–72, 2010.

[84] L. Mai, H. Le, Y. Niu, and F. Liu. Rule of Thirds Detection from Photograph. In *Multimedia (ISM), 2011 IEEE International Symposium on*, pages 91–96, 2011.

[85] S. Marat. *Modèles de saillance visuelle par fusion d'informations sur la luminance, le mouvement et les visages pour la prédiction de mouvements oculaires lors de l'exploration de vidéos*. PhD thesis, université de grenoble, February 2010.

[86] S. Marat, T. Ho Phuoc, L. Granjon, N. Guyader, D. Pellerin, and A. Guérin-Dugué. Spatio-temporal saliency model to predict eye movements in video free viewing. In *16th European Signal Processing Conference (EUSIPCO-2008)*, pages 1–5, Lausanne, Switzerland, August 2008.

[87] S. Marat, T. Phuoc, L. Granjon, N. Guyader, D. Pellerin, and A. Guerin-Dugue. Modelling spatiotemporal saliency to predict gaze direction for short videos. *International Journal of Computer Vision*, (82):231–243, 2009.

[88] M. Marszałek, I. Laptev, and C. Schmid. Actions in Context. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 2929–2936, June.

[89] S. Mathe and C. Sminchisescu. Actions in the Eye: Dynamic Gaze Datasets and Learnt Saliency Models for Visual Recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 37(7):1408–1424, July 2015.

[90] G. Mesnil, Y. Dauphin, X. Glorot, S. Rifai, Y. Bengio, I. J. Goodfellow, E. Lavoie, X. Muller, G. Desjardins, D. Warde-Farley, P. Vincent, A. Courville, and J. Bergstra. Unsupervised and Transfer Learning Challenge: a Deep Learning approach. In *JMLR W& CP: Proceedings of the Unsupervised and Transfer Learning challenge and workshop*, volume 27, pages 97–110, 2012.

[91] O. Meur, P. L. Callet, and D. Barba. Predicting visual fixations on video based on low-level visual features. *Vision Research*, (47):2483–2498, 2007.

[92] R. Monroy, S. Lutz, T. Chalasani, and A. Smolic. Salnet360: Saliency maps for omni-directional images with CNN. *CoRR*, abs/1709.06505, 2017.

[93] C. Morand, J. Benois-Pineau, and J. P. Domenger. Scalable Object-Based Indexing of HD Videos: A JPEG2000- Oriented solution. working paper or preprint, December 2008.

[94] F. Murabito, C. Spampinato, S. Palazzo, K. Pogorelov, and M. Riegler. Top-Down Saliency Detection Driven by Visual Classification, 2017. arXiv:1709.05307v2.

[95] R. Nakayama, I? Motoyoshi, and T. Sato. The roles of non-retinotopic motions in visual search. *Frontiers in Psychology*, 7(840), 2016.

[96] Y. Nesterov. A method of solving a convex programming problem with convergence rate $O\left(1/k^2\right)$. In *Soviet Mathematics Doklady*, volume 27, pages 372–376, 1983.

[97] D. F. Nettleton, A. Orriols-Puig, and A. Fornells. A study of the effect of different types of noise on the precision of supervised learning techniques. *Artificial Intelligence Review*, 33(4):275–306, 2010.

[98] A. Y. Ng, J. Ngiam, C. Y. Foo, Y. Mai, and C. Suen. UFLDL deep learning tutorial. Technical report, Stanford University, 2013.

[99] D. Noton and L. Stark. Scanpaths in saccadic eye movements while viewing and recognizing patterns. *Vision Research*, 11(9):929–IN8, 1971.

[100] A. Palazzi, D. Abati, S. Calderara, F. Solera, and R. Cucchiara. Predicting the Driver's Focus of Attention: the DR(eye)VE Project. *CoRR*, abs/1705.03854, 2017.

[101] J. Pan and X. Giró i Nieto. End-to-end Convolutional Network for Saliency Prediction. *IEEE Conference on Computer Vision and Pattern Recognition*, 1507.01422, 2015.

[102] J. Pan and Giró i Nieto, X. End-to-end Convolutional Network for Saliency Prediction. *CoRR*, abs/1507.01422, 2015.

[103] E. Peli. Contrast in complex images. *J. Opt. Soc. Am. A*, 7(10):2032–2040, Oct 1990.

[104] Y. Pinto, A. R. van der Leij, I. G. Sligte, V.A. F. Lamme, and H. S. Scholte. Bottom-up and top-down attention are independent. *Journal of Vision*, 13(3):16, 2013.

[105] B.T. Polyak. *Introduction to Optimization (Translations Series in Mathematics and Engineering).* Optimization Software, 1987.

[106] W. H. Press, S. A. Teukolsky, W. T. Vetterling, and B. P. Flannery. *Numerical Recipes in C (2Nd Ed.): The Art of Scientific Computing.* Cambridge University Press, New York, NY, USA, 1992.

[107] D. Purves, G. J. Augustine, D. Fitzpatrick, L. C. Katz, A. S. LaMantia, J. O. McNamara, and S. M. Williams. *Neuroscience, 2nd edition.* Sinauer Associates, 2001.

[108] V. Ramanishka, A. Das, J. Zhang, and K. Saenko. Top-down Visual Saliency Guided by Captions. *CoRR*, abs/1612.07360, 2016.

[109] O. M. Ranzato, C. Poultney, S. Chopra, and Y. LeCun. Efficient Learning of Sparse Representations with an Energy-Based Model. In P. B. Schölkopf, J. C. Platt, and T. Hoffman, editors, *Advances in Neural Information Processing Systems 19*, pages 1137–1144. MIT Press, 2007.

[110] S. E. Reed, H. Lee, D. Anguelov, C. Szegdy, D. Erhan, and A. Rabinovich. Training Deep Neural Networks on Noisy Labels with Bootstrapping. *CoRR*, abs/1412.6596, 2014.

[111] F. Rosenblatt. The Perceptron: A Probabilistic Model for Information Storage and Organization in The Brain. *Psychological Review*, pages 65–386, 1958.

[112] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei. Imagenet large scale

visual recognition challenge. *International Journal of Computer Vision (IJCV)*, 115(3):211–252, 2015.

[113] H. J. Seo and P. Milanfar. Static and space-time visual saliency detection by self-resemblance. *Journal of Vision*, (9(12):15):1–27, 2009.

[114] E. Shelhamer, J. Long, and T. Darrell. Fully Convolutional Networks for Semantic Segmentation. *IEEE Conference on Computer Vision and Pattern Recognition*, 1605.06211, 2015.

[115] C. Shen and Q. Zhao. Learning to Predict Eye Fixations for Semantic Contents Using Multi-layer Sparse Network. *Neurocomputing*, 138:61–68, 2014.

[116] D. Simon, S. Sridharan, S. Sah, R. Ptucha, C. Kanan, and R. Bailey. Automatic Scanpath Generation with Deep Recurrent Neural Networks. In *Proceedings of the ACM Symposium on Applied Perception*, SAP '16, pages 130–130, New York, NY, USA, 2016. ACM.

[117] K. Simonyan, A. Vedaldi, and A. Zisserman. Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps. *CoRR*, abs/1312.6034, 2013.

[118] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *CoRR*, abs/1409.1556, 2014.

[119] H. Snellen. Probebuchstaben zur bestimmung der sehscharfe. P.w. van de weijer, utrecht, 1862.

[120] I. Sutskever, J. Martens, G. Dahl, and G. Hinton. On the importance of initialization and momentum in deep learning. In Sanjoy Dasgupta and David McAllester, editors, *Proceedings of the 30th International Conference on Machine Learning*, volume 28 of *Proceedings of Machine Learning Research*, pages 1139–1147, Atlanta, Georgia, USA, 17–19 Jun 2013. PMLR.

[121] W. Szegedy, C.and Liu, Y. Jia, P. Sermanet, S. E. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. Going deeper with convolutions. *CoRR*, abs/1409.4842, 2014.

[122] H. R. Tavakoli, A. Borji, J. Laaksonen, and E. Rahtu. Exploiting inter-image similarity and ensemble of extreme learners for fixation prediction using deep features. *CoRR*, abs/1610.06449, 2016.

[123] F. Tison and G. Chene. Les Yeux L'Ont: anomalies des saccades oculaires à la phase prodromale de la maladie d'Alzheimer ACRONYME : LYLO. *PROTOCOLE DE RECHERCHE BIOMEDICALE Version nÂ° 3.0 du 09/10/2013*, 2013.

[124] A. Torralba, R. Fergus, and W. Freeman. 80 million tiny images: A large dataset for non-parametric object and scene recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, (30):1958–1970, 2008.

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

[126] P. Tseng, I. G. M. Cameron, G. Pari, J. N. Reynolds, D. P. Munoz, and L. Itti. High-throughput classification of clinical populations from natural viewing eye movements. *Journal of Neurology*, 260:275–284, Jan 2013.

[127] P. Tseng, A. Paolozza, D. P. Munoz, J. N. Reynolds, and L. Itti. Deep learning on natural viewing behaviors to differentiate children with fetal alcohol spectrum disorder. In *The 14th International Conference on Intelligent Data Engineering and Automated Learning (IDEAL 2013), Hefei, China (LNCS 8206)*, pages 178–185, Oct 2013.

[128] J. M. G. Tsui and C. C. Pack. Contrast sensitivity of MT receptive field centers and surrounds. *Journal of Neurophysiology*, 106(4):1888–1900, 2011.

[129] Vladimir V. Principles of Risk Minimization for Learning Theory. In John E. Moody, Stephen Jose Hanson, and Richard Lippmann, editors, *NIPS*, pages 831–838. Morgan Kaufmann, 1991.

[130] E. Vig, M. Dorr, and D. Cox. Large-Scale Optimization of Hierarchical Features for Saliency Prediction in Natural Images . In *Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition*, CVPR '14, pages 2798–2805, 2014.

[131] D. S. Wooding. Eye movements of large populations: II. Deriving regions of interest, coverage, and similarity using fixation maps. *Behavior Research Methods, Instruments, & Computers*, 34(4):518–528, November 2002.

[132] T. Xiao, T. Xia, Y. Yang, Ch. Huang, and X. Wang. Learning From Massive Noisy Labeled Data for Image Classification. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2015.

[133] A.L. Yarbus. *Eye movements and vision.* Plenum Press, New York, 1967.

[134] J. Yosinski, J. Clune, Y. Bengio, and H. Lipson. How transferable are features in deep neural networks? In Z. Ghahramani, M. Welling, C. Cortes, N.d. Lawrence, and K.q. Weinberger, editors, *Advances in Neural Information Processing Systems 27*, pages 3320–3328. Curran Associates, Inc., 2014.

[135] L. Yuetan, K. Shu, W. Donghui, and Z. Yueting. Saliency detection within a deep convolutional architecture. 2014.

[136] M. D. Zeiler and R. Fergus. Visualizing and Understanding Convolutional Networks. In David Fleet, Tomas Pajdla, Bernt Schiele, and Tinne Tuytelaars, editors, *Computer Vision - ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part I*, pages 818–833, Cham, 2014. Springer International Publishing.

[137] D. Zhang, H. Fu, J. Han, and F. Wu. A Review of Co-saliency Detection Technique: Fundamentals, Applications, and Challenges. *CoRR*, abs/1604.07090, 2016.

[138] L. Zhang, Y. Shen, and H. Li. VSI: A Visual Saliency-Induced Index for Perceptual Image Quality Assessment. *IEEE Transactions on Image Processing*, 23(10):4270–4281, 2014.

[139] J. T. Zhou, S. J. Pan, I. W. Tsang, and Y. Yan. Hybrid Heterogeneous Transfer Learning through Deep Learning. In *Proceedings of the Twenty-Eighth AAAI Conference on Artificial Intelligence*, pages 2213–2220, July 2014.

[140] O. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba. Learning deep features for discriminative localization. *IEEE Conference on Computer Vision and Pattern Recognition*, 1512.04150, 2015.

[141] X. Zhu and X. Wu. Class Noise vs. Attribute Noise: A Quantitative Study. *Artificial Intelligence Review*, 22(3):177–210, 2004.

Publications

International Journal

– S. Chaabouni, J. Benois-Pineau, F. Tison, Ch. Ben Amar, A. Zemmari "Prediction of visual attention with Deep CNN on artificially degraded videos for studies of attention of Patients with Dementia.", Multimed Tools Appl, 2017

Book Chapter

– S. Chaabouni, J. Benois-Pineau, Ch. Ben Amar, A. Zemmari "Deep Saliency: Prediction of Interestingness In Video With CNN.", Chapter book "Visual content indexing and retrieval with psycho-visual models" in Springer, 2017.

International Conferences with Proceedings

– S. Chaabouni, J. Benois-Pineau, Ch. Ben Amar, Transfer learning with deep networks for saliency prediction in natural video, ICIP, 2016.

– S. Chaabouni, J. Benois-Pineau, O. Hadar, Prediction of visual saliency in video with deep CNNs, Proc.SPIE 9971, Applications of Digital Image Processing XXXIX, 99711Q, 2016.

– S. Chaabouni, J. Benois-Pineau, F. Tison, Ch. Ben Amar, Prediction of visual attention with deep CNN for studies of neurodegenerative diseases, CBMI, 2016.

National Conferences with Proceedings

– S. Chaabouni, J. Benois-Pineau, Ch. Ben Amar, Sensitivity of Deep CNNs to noise in training Data in the problem of visual saliency prediction, ERIS, 2016.

Other

– S. Chaabouni, J. Benois-Pineau, O. Hadar, Ch. Ben Amar, Deep network for saliency prediction in natural video, http://arxiv.org/abs/1604.08010, 2016.

– S. Chaabouni, J. Benois-Pineau, Ch. Ben Amar "ChaboNet : Design of a deep CNN for prediction of visual saliency in natural video", in Preparation.

Appendix A

LYLO protocol

« Les Yeux L'Ont » : anomalies des saccades oculaires à la phase prodromale de la maladie d'Alzheimer
ACRONYME : LYLO

Code promoteur : CHUBX 2011/22

PROTOCOLE DE RECHERCHE BIOMEDICALE
Version n°3.0 du 09/10/2013

Numéro ID-RCB : 2011-A01574-37
Numéro d'enregistrement dans Clinicaltrials.gov : NCT01630525

Cette recherche biomédicale a obtenu le financement du PHRC national 2011

Promoteur :
Centre Hospitalier Universitaire de Bordeaux
12, rue Dubernat
33400 Talence
France

Investigateur coordinateur :
Pr François TISON,
Service de Neurologie,
GH Sud, CHU de Bordeaux, Hôpital Haut-Lévêque
Av. Magellan, 33604 Pessac Cedex, France
Tél : 05 57 65 64 20 – Fax : 05 57 65 68 15
Courriel : francois.tison@chu-bordeaux.fr

Centre de Méthodologie et de Gestion des données :
Pr Geneviève CHENE
Unité de Soutien Méthodologique à la Recherche Clinique et Epidémiologique du CHU de Bordeaux, 146 rue Léo-Saignat, case n°11
33076 Bordeaux Cedex
Tel : 05 57 57 11 29- Fax : 05 57 57 15 78
Courriel : genevieve.chene@isped.u-bordeaux2.fr

Unité de sécurité et de vigilance de la Recherche Clinique :
Direction de la recherche clinique et de l'innovation
12, rue Dubernat
33404 Talence Cedex
Tél. : 05 57 82 08 34 - Fax : 05.57.82.12.62
Courriel : vigilance.essaiscliniques@chu-bordeaux.fr

Ce protocole a été conçu et rédigé à partir de la version 1.0 du 14/11/2008 du protocole-type de la DIRC Sud-Ouest Outre Mer

Historique des mises à jour du protocole

Version	Date	Raison de la mise à jour
1.0	07/12/2011	Soumission CPP et AFSSAPS
1.1	03/01/2012	Complément d'information CPP
2.0	08/08/2012	Amendement n°1
3.0	09/10/2013	Modification substantielle N°2

PRINCIPAUX CORRESPONDANTS

Investigateur coordonnateur

Pr François Tison
Service de neurologie,
Hôpital du Haut-Lévêque
Avenue Magellan
33604 Pessac Cedex, FRANCE
Tél.: 05 57 65 64 61- Fax : 05 57 65 66 97
Courriel : françois.tison@chu-bordeaux.fr

Attachées de Recherche Clinique coordonnatrices

Sandrine Dupouy
Service de Neurologie- Hôpital du Ht-Lévêque
Avenue Magellan
33604 Pessac Cedex, FRANCE
Tél. : 05 57 65 69 14 - Fax : 05 57 65 66 97
Courriel : sandrine.dupouy@chu-bordeaux.fr

Unité de sécurité et de vigilance de la Recherche Clinique :

Direction de la recherche clinique et de l'innovation
12, rue Dubernat
33404 Talence Cedex
Tél. : 05 57 82 08 34 - Fax : 05.57.82.12.62
Courriel : vigilance.essaiscliniques@chu-bordeaux.fr

Autres spécialités

Prestataire de service, logiciels à façon
Serge Kinkingnéhun, PhD
e(ye)BRAIN A computer vision of the medical world
3 bis, rue Maurice Grandcoing
94200 Ivry-sur-Seine
01 78 57 35 76
eye.brain@gmail.com

Promoteur

Centre Hospitalier Universitaire de Bordeaux
12, rue Dubernat
33400 Talence
FRANCE

Responsable de la recherche au niveau du promoteur

Joaquin Martinez - Directeur de la Recherche Clinique et de l'Innovation
Julie Boussuge-Roze - Responsable du département Promotion Institutionnelle
Tél : 05 57 82 03 13 – Fax : 05 56 79 49 26
Courriel : julie.boussuge-roze@chu-bordeaux.fr

Responsable d'Etudes Cliniques

Laetitia LACAZE-BUZY
Tél. : 05 57 82 11 34 - Fax : 05 56 79 49 26
Courriel : laetitia.lacaze-buzy@chu-bordeaux.fr

Coordination méthodologique

Pr CHENE Geneviève
Tel : 05 57 57 11 29 – Fax : 05 57 57 15 78
Courriel :genevieve.chene@isped.u-bordeaux2.fr

Statisticienne

Julien ASSELINEAU
Tel : 05 57 57 45 14 – Fax : 05 57 57 15 78
Courriel : julien.asselineau@isped.u-bordeaux2.fr

Analyste Programmeur

Guillaume Dupouy
Tel : 05 57 57 95 79 – Fax : 05 57 57 15 78
Courriel :helene.savel@isped.u-bordeaux2.fr

SOMMAIRE

PAGE DE SIGNATURE DU PROTOCOLE

« Les yeux l'ont : Anomalies des saccades oculaires à la phase prodromale de la maladie d'Alzheimer »

LYLO

Code promoteur : CHUBX 2011/22

Promoteur

Centre Hospitalier Universitaire de Bordeaux
12, rue Dubernat
33400 Talence

à Talence, le 09/10/2013

Le Directeur Général du CHU de Bordeaux

Philippe VIGOUROUX

Et par délégation, le Directeur de la Recherche Clinique et de l'Innovation,

Joaquin MARTINEZ

Investigateur coordonnateur

Pr François TISON,
Service de Neurologie,
GH Sud, CHU de Bordeaux
Hôpital Haut-Lévêque
Av. Magellan, 33604 Pessac Cedex
Tél : 05 57 65 64 20
Courriel : francois.tison@chu-bordeaux.fr

à Pessac, le 09/10/2013

Pr François TISON

LISTE DES ABRÉVIATIONS

A DAS-cog :	Alzheimer Disease Assessment Scale-Cognitive
AFSSAPS :	Agence Française de Sécurité Sanitaire des Produits de Santé
ANSM :	Agence nationale de sécurité du médicament
AP-HP :	Assistance Publique-Hôpitaux de Paris
ARC:	Attaché de Recherche Clinique
CCTIRS :	Comité consultatif sur le traitement de l'information en matière de recherche dans le domaine de la santé
CDR :	Clinical Dementia Rating Scale
CMRR :	Centres Mémoire de Recherche et de Ressources
CNIL :	Commission Nationale de l'Informatique et des Libertés
CPP :	Comité de Protection des Personnes
DMS:	Delayed Matching to Sample
DO 80:	Dénomination Orale 80
DSST :	Digit Symbol Substitution Task
EI:	Evénement indésirable
EIG:	Evénement indésirable grave
FO:	Fond d'œil
GDS :	Geriatric Depression Scale
IADL:	Instrumental Activities of Daily Living
IRM :	Imagerie par Résonance Magnétique
LCR :	Liquide Céphalo-rachidien
MA :	Maladie d'Alzheimer
MAP :	Maladie d'Alzheimer Prodromale
MCI:	Mild Cognitive Impairment
MMSE :	Mini Mental State Examination
MTA :	Medial Temporal Lobe Atrophy
OCT :	Optical Coherence Tomography
RL/RI :	Rappel libre / Rappel indicé
TMT:	Trail Making Test
USMR :	Unité de Soutien Méthodologique à la Recherche Clinique et Epidémiologique
WAIS :	Wechsler Adult Intelligence Scale

RESUME

PROMOTEUR	CHU de Bordeaux
INVESTIGATEUR COORDINATEUR	Pr François TISON, Pôle Neurosciences Cliniques, GH Sud, CHU de Bordeaux
TITRE	LYLO « Les Yeux L'Ont » : anomalies des saccades oculaires à la phase prodromale de la maladie d'Alzheimer
JUSTIFICATION / CONTEXTE	La maladie d'Alzheimer (MA) n'est actuellement diagnostiquée qu'au stade démentiel, ainsi le concept de MA prodromale (MAP) a été proposé pour identifier la maladie à un stade plus précoce, pré-démentiel (*Dubois et al. Lancet Neurol 2010 ; 9 :118-27*). Les fonctions exécutives (frontales), à l'interface entre cognition et action, sont altérées plus de 10 ans avant l'installation du syndrome démentiel (*Amieva et al. Ann Neurol. 2008;64:492-8*). Des tests comme la substitution chiffres/symboles (Digit Symbol Substitution Task : DSST) ayant une composante attentionnelle, de vitesse d'exécution et d'intégration visuospatiale sont ainsi très sensibles à cette altération précoce. L'exécution des mouvements oculaires rapides dirigés vers un but ou saccades volontaires se situe précisément à l'interface des processus fronto-pariétaux d'attention, de prise de décision et de mémoire de travail spatiale. Des défauts d'exécution, d'inhibition des saccades et de stratégie d'exploration visuelle ont été mis en évidence aux stades avancés de MA (*Garbutt et al. Brain. 2008;131:1268-81*). Cependant, il n'existe pas d'étude réalisée chez des patients à des stades plus précoces, alors que l'exécution des saccades volontaires implique des structures touchées relativement tôt au cours du processus physiopathologique de la MA (ex : noyau caudé, pré-cuneus).
OBJECTIFS	*Objectif principal de la recherche :* Comparer la distribution des paramètres d'exécution des saccades (latence, vitesse, précision, erreurs) dans des taches de pro- saccades horizontales et verticales, d'inhibition (anti-saccades), de prédiction et de décision spatiale (selon *Monsimann et al. Brain 2005 ;128:1267-1276*) dans la MAP comparée à des MA probables légères à modérées et à des sujets contrôles (C) appariés pour l'âge. *Objectif secondaires :* - Mesurer les altérations des stratégies d'exploration visuelle (test de détection d'items-*Rösler et al. Cortex 2005 ;41 :512-519* et exploration de visages, images incongrues, selon *Daffner et al. Neurology 1992;42 :320-328 et Laughland et al. Biol Psychiatry*

	2002 ;52 : 338-348). - Etudier l'association entre l'altération des saccades ou les stratégies d'exploration visuelle et les performances cognitives en particulier celles obtenues au DSST. - Comparer le nombre moyen de points de fixation dans les régions dégradées et des cartes d'attentions visuelles induites
SCHEMA DE LA RECHERCHE	Etude multicentrique transversale physiopathologique de type « cas-témoins » comprenant 3 groupes de patients : MAP, MA et contrôles.
CRITERES D'INCLUSION	**Tous groupes :** - Age > 60 ans - Sujets ayant donné leur consentement éclairé et affiliés à un régime de Sécurité Sociale. Le consentement peut être recueilli par le représentant du patient le cas échéant. **Groupe A : Maladie d'Alzheimer prodromale (MAP) ou pré-démentielle** (critères Dubois et Albert Lancet Neurol. 2004 ; 3:246-8 et revus Dubois et al. Lancet Neurol 2010 ; 9 :118-27). - Plainte mnésique rapportée par le patient ou la famille - Activités complexes de la vie quotidienne (IADL), normales ou légèrement altérées (seulement le premier niveau atteint) - Syndrome amnésique de type hippocampique défini par un rappel libre très faible malgré un encodage adéquat, diminution du rappel total en raison de l'absence d'effet de l'indiçage (RL/RI-16items = rappel libre total inférieur à 17 ou rappel total inférieur à 40) ou trouble de la reconnaissance ; intrusions nombreuses) - CDR (Clinical Dementia Rating Scale) ≤ 0,5 - Persistance des troubles mnésiques à une évaluation ultérieure (> 3 mois) - Absence de démence totalement développée (MMSE ≥ 24), ou suivant l'appréciation du clinicien - Exclusion d'autres maladies pouvant causer un trouble cognitif léger (MCI), - IRM cérébrale montrant une atrophie médio-temporale/hippocampique ou hypométabolisme temporo-pariétal en TEP ou SPECT ou si LCR disponible (non obligatoire) ratio bétaA42/tau anormal. **Groupe B : Maladies d'Alzheimer typiques légères à modérées** - Critères diagnostiques NINDS-ADRDA actualisés selon *Dubois et al. Lancet Neurol 2007;6 :734-46 et 2010 ; 9 :118-27.* - MMSE ≥ 20 **Groupe C : Sujets contrôles**

	- Pas de plainte mnésique ou cognitive - MMSE ≥ 24 si bas niveau d'étude (CEP -) - MMSE ≥ 26 si haut niveau d'étude (CEP ou +)
CRITERES DE NON INCLUSION	**Tous groupes :** - Sujet ne lisant pas P8 de prés avec correction - Troubles de l'oculomotricité et strabismes à l'appréciation du clinicien - Dépression (diagnostic clinique ou score GDS ≥ 10) - Sujet ne pouvant donner son consentement éclairé **Contrôles :** - Plainte mnésique ou autre plainte cognitive significative. - Anomalie au bilan neuropsychologique à V0 laissant supposer une dégradation des fonctions cognitives
PROCEDURES DE LA RECHERCHE	**V0, Visite de pré-inclusion/inclusion (durée 2 hres)** - Critères d'inclusion/exclusion - Consentement éclairé - Bilan neuropsychologique si date de plus de 3 mois et pour groupe contrôle : MMSE (version Greco), RL/RI-16items (Van der Linden 2003), Test de reconnaissance visuelle (DMS48), fluence verbale littérale et catégorielle (Thurstone et Thurstone 1964), TMT A et B (Reitan 1956), Test des codes (DSST, Weschler 1997), Clinical Dementia Rating Scale (Hughes 1982), ,dénomination d'images DO80 (Deloche et Hannequin 1997), Test des similitudes et Digit Span de la WAIS (Weschler 1997), anxiété et dépression (GDS) et activités de la vie quotidienne (ADL-Katz et IADL-Lawton) si date de plus de 3 mois (bilan effectué à titre diagnostique) **V1, Visite d'étude (durée 1 heure 30), dans le mois suivant la V0** - MMSE et IADL. - Oculométrie automatisée non-invasive 45 minutes avec pauses : poursuite, saccades horizontales, saccades verticales, anti-saccades, prédiction, décision spatiale (*Monsiman et al. Brain 2005 ; 128:1267-1276*), détection d'Items (*Rösler et al. Cortex 2005 ; 41 :512*-519) et curiosité avec exploration de visages, images incongrues, selon *Daffner et al. Neurology 1992 ; 42 :320-328 et Loughland et al. Biol Psychiatry 2002 ; 52 : 338-348*) et de vidéos naturelles dégradées et non dégradées.
CRITERES DE JUGEMENT	Critère de jugement principal de la recherche : paramètres d'exécution des saccades - La latence moyenne de déclenchement (msec). - Vitesse moyenne (°/msec) et vitesse maximale. - Précision ou gain moyen (ratio composante A1/A2). - Le pourcentage moyen d'erreurs et pourcentages d'erreurs corrigées. - Pourcentage moyen de prédiction.

	Critères de jugement secondaires : - Scores aux différents tests du bilan neuropsychologique. - Variables définies lors des tests de détection d'items et d'exploration des visages, d'images incongrues (nombre de fixation, durée des fixations, erreurs). - Nombre moyen de points de fixations dans les régions dégradées et des cartes d'attentions visuelles induites
NOMBRE DE SUJETS PREVUS	**30 MAP** (Groupe A), **30 MA** (Groupe B) et **30 contrôles** (Groupe C, appariés ± 5ans avec groupe A) soit **90 patients au total.**
NOMBRE PREVU DE CENTRES	3 (CHU de Bordeaux, CHU de Lyon, Hôpital de la Timone-Marseille)
DUREE DE LA RECHERCHE	Durée de la période d'inclusion et d'étude = 24 mois Durée de participation de chaque patient = 1 mois max Durée totale de la recherche = 30 mois incluant période préparatoire (3 mois), période d'étude (24 mois) et période analyse/rapport final (3 mois).
ANALYSE STATISTIQUE DES DONNEES	L'objectif principal est de comparer la distribution des paramètres d'exécution des saccades, dans un premier temps entre le groupe MAP et le groupe contrôle puis entre le groupe MAP et le groupe MA. Les paramètres étant des variables quantitatives, ils seront décrits en termes d'effectif, moyenne, écart-type et intervalle de confiance à 95% de la moyenne, médiane, minimum, maximum, Q1 et Q3. Les comparaisons 2 à 2 entre les groupes des différents paramètres utiliseront le test de Student pour des données appariées (comparaison MAP/contrôles) ou non appariées (comparaison MAP/MA) ou le test non paramétrique de Wilcoxon pour données appariées (comparaison MA/contrôles) ou non appariées (comparaison MAP/MA) si l'hypothèse de normalité n'est pas vérifiée.
RETOMBEES ATTENDUES	Contribution à la définition diagnostique multimodale de MA prodromale (biomarqueur). Compréhension de la physiopathologie de l'altération précoce de tests comme le DSST (composante oculomotrice et cognitive). Nouvelle variable d'essai thérapeutique dans les phases précoces de MA.

ABSTRACT

SPONSOR	CHU de Bordeaux (Bordeaux University Hospital)
COORDINATOR INVESTIGATOR	Pr François Tison, Pôle Neurosciences Cliniques, GH Sud, CHU de Bordeaux
ACRONYM and TITLE	LYLO « The eyes have it » : ocular saccade abnormalities in prodromal Alzheimer's disease
BACKGROUND/PURPOSE	Alzheimer's disease (AD) has a prolonged prodromal phase (prodromal PAD, *Dubois B et al. Lancet Neurol 2010; 9 :118-27*) before the stage of dementia. Subtle executive cognitive function deficits can be detected at this early pre-dementia phase, more than 10 years before dementia (*Amieva et al. Ann Neurol. 2008; 64:492-8*). Among them, the digit symbol substitution task (DSST) has been shown to be altered very early, up to 13 years before dementia. This test, as many others executive function tests, requires a fine control of visuomotor coordination. Like executive functions, eye movements, particularly voluntary-guided saccades, are under the control of the frontal lobe and fronto-parietal networks. Previous studies have shown a deterioration of voluntary saccades in AD using various paradigms *(Garbutt et al. Brain. 2008; 131:1268-81).* There are no data in PAD, although the pathological process of the disease affects very early brain structures implicated in saccades execution (eg. caudate nucleus and pre-cuneus).
OUTCOMES	*Primary outcome:* To demonstrate the alteration of saccade execution parameters (latency, velocity, precision, errors) during pro and anti-saccades, spatial decision and prediction tasks (according to *Monsimann et al. Brain 2005; 128:1267-1276*) in prodromal AD compared to mild to moderate AD and aged-matched controls. *Secondary outcomes*: - To demonstrate alterations in visual search strategies and curiosity (items detection tasks -*Rösler et al. Cortex 2005; 41:512*-519- and faces and non congruent images exploration according to *Daffner et al. Neurology 1992; 42 :320-328 and Loughland et al. Biol Psychiatry 2002; 52:338-348*). - To correlate saccades and visual search strategy parameters with cognitive performances, particularly that of the DSST. - Comparison of the average point fixation in degraded areas and of visual attention induced cards
STUDY DESIGN	Multicentre "Case-control" pathophysiological transversal study studying 3 groups: PAD, AD and controls.

ELIGIBILITY CRITERIA: INCLUSION CRITERIA	**All patient groups:** - Age >60 years - Written informed consent - Subjects affiliated to Social Security **Group A: Prodromal AD (PAD)** (criteria Dubois and Albert Lancet Neurol. 2004; 3:246-8, revised Dubois B et al. Lancet Neurol 2010; 9:118-27). - Memory complaints. - Normal or slight restriction of IADL. - "hippocampal-type" amnesic syndrome defined by poor free recall despite adequate (and controlled) encoding, decreased total recall because of insufficient effect of cuing or impaired recognition, numerous intrusions (RL/RI-16items). - CDR (Clinical Dementia Rating Scale) ≤ 0,5 - Persistence of memory changes at a subsequent assessment (>3 months). - Absence of global cognitive deterioration (MMSE ≥24). - Exclusion of other disorders that may cause mild cognitive impairment with adequate tests - Cerebral MRI showing medio temporal/hippocampal atrophy or PET/SPECT showing hypoperfusion in temporo-parietal regions or if available (non mandatory) abnormal CSF betaA42/tau ratio. **Group B: Typical AD (mild to moderate**) - NINDS-ADRDA diagnosis criteria, revised according to *Dubois et al. Lancet Neurol 2007;6 : 734-46 and 2010 ;9 :118-27* - MMSE ≥ 20. **Group C: Control subjects** - No memory or other significant cognitive complain. - MMSE ≥ 24 if low-level study (CEP -) - MMSE ≥ 26 if high-level study (CEP ou +)
ELIGIBILITY CRITERIA: EXCLUSION CRITERIA	**All groups :** - Subject that can not read P8 with correction of near - Oculomotor deficit or strabismus at the opinion of the clinician - Depression (clinical diagnosis or GDS score ≥ 10) - Subjects unable to give their informed consent. **Controls :** - Memory or any other significant cognitive complain. - Abnormalities at inclusion (V0) neuropsychology testing suggestive of a cognitive deficit.
RESEARCH PROCEDURE/INTERVENTIONS	**V0, pre-Inclusion/inclusion visit:** 2 hours duration - Inclusion/exclusion criteria. - Written informed consent. - Neuropsychological assessment for subjects with tests > 3 mo and control group : MMSE (Greco), RL/RI-16 items (Van der Linden 2003), visual retention test (DMS48), verbal fluency (Thurstone et Thurstone 1964), TMT A and B (Reitan 1956), DSST (Weschler 1997), Clinical Dementia Rating Scale

	(Hughes 1982), image naming DO80 (Deloche et Hannequin 1997), Similarities and Digit Span subscores of the WAIS (Weschler 1997), Anxiety and Depression (GDS), activities of daily living (ADL-Katz and IADL-Lawton) if not already performed during the past 3 months for diagnosis purpose. **V1, study visit (1.5 hour duration)**, within a month after V0 - MMSE and IADL. - Automated non-invasive oculometry : 45 minutes with rest periods : horizontal and vertical pro- and anti-saccades, prediction, spatial decision (*Monsiman et al. Brain 2005,128:1267-127*, items detection (*Rösler et al. Cortex 2005 ;41 :512*-519), exploration/curiosity of non congruent images and faces according to *Daffner et al. Neurology 1992 ;42 :320-328 and Loughland et al. Biol Psychiatry 2002 ;52 : 338-348*) and degraded and not degraded natural videos.
VARIABLES	*Primary variables*: saccades execution parameters - Mean latency (msec). - Mean velocity (°/msec) and maximal velocity. - Accuracy or mean gain (ratio A1/A2). - Mean percent of errors and corrected errors. - Mean percent of prediction. *Secondary variables :* - Neuropsychology tests scores. - Pre-defined variables on visual exploration tasks (fixation number and durations, errors). - Number of point fixation in degraded areas and of visual attention induced cards
NUMBER OF SUBJECTS	**30 PAD** (Group A), **30 typical AD** (Group B) and **30 age-matched** (to Group A) **controls** (Group C, matched by age to group A ± 5 years), thus a total of **90 subjects**.
INVESTIGATION CENTERS	**3** (CHU de Bordeaux, CHU de Lyon, Hôpital de la Timone-Marseille)
RESEARCH CALENDAR	Inclusion period duration = 24 months Subject participation duration = 1 month max Total research duration = 30 months including research setting (3 months), research period (24 months) and data analysis/final report (3 months).
STATISTICAL ANALYSIS	The primary objective is to compare the distribution of saccade execution parameters, firstly between PAD and controls then between PAD and AD. Quantitative variables will be described as number of subjects, mean, and SD, 95% confidence interval to the mean, median and range. Comparisons 2 by 2 between groups of the various parameters will use a Student test (paired for PAD/controls) or a non parametric Wilcoxon test (paired for PAD/controls) if the normal distribution is not verified.
RESEARCH OUTCOMES	Contribution to the definition and early diagnosis of prodromal AD (biomarker). Pathophysiology of the early deterioration of the DSST (respective contribution to the oculomotor and cognitive component). New variables for early drug trials in AD.

JUSTIFICATION SCIENTIFIQUE ET DESCRIPTION GÉNÉRALE

1.1. ETAT ACTUEL DES CONNAISSANCES

La maladie d'Alzheimer (MA) n'est habituellement définie et le plus souvent diagnostiquée qu'au stade de démence (défini par l'altération des activités de la vie quotidienne), correspondant à des lésions neuropathologies avancées, ne pouvant laisser espérer qu'une action thérapeutique tardive et limitée dans l'histoire naturelle de la maladie (Dubois et al, 2007). Les concepts de maladie d'Alzheimer prodromale (MAP) ou pré-démentielle ont été proposés pour définir une phase plus précoce de la maladie amenant les patients à consulter pour une plainte mnésique d'apparition progressive sans altération des activités de la vie quotidienne (Dubois et Albert, 2004, Dubois et al, 2007, Dubois et al ; 2010). La définition de ce stade prodromal pré-démentiel de la MA est implémentée par la valeur prédictive de marqueurs multimodaux : cliniques (syndrome amnésique de type hippocampique), d'imagerie structurelle et/ou fonctionnelle (ex : atrophie hippocampique en IRM, et imagerie amyloïde 11C-PIB PET) et biomarqueurs (Abéta42/tau dans le LCR) (Dubois et al. 2007, Okello et al. 2009 ; Petrie et al. 2009, Dubois et al ; 2010). Les altérations neuropsychologiques précoces pré-démentielles de la MA semblent également aller au-delà des fonctions mnésiques temporo-hippocampiques caractéristiques de la maladie, ainsi les fonctions exécutives supportées par les boucles sous-cortico-frontales et le cortex frontal, seraient encore plus précocement altérées (Fabrigoule et al. 1998, Amieva et al 2008, Okello et al. 2009). Des tests comme la substitution chiffres/symboles (*Digit Symbol Substitution Task :DSST* ou Test des codes de Wechsler) ayant une composante attentionnelle, de vitesse d'exécution et d'intégration visuospatiale sont ainsi très sensibles à cette altération précoce (Amieva et al 2010, en préparation). L'altération de ces fonctions cognitives est parfaitement connue et caractérisée dans les maladies affectant la motricité extrapyramidale et les noyaux gris centraux comme dans les syndromes parkinsoniens (Rivaud-Péchoux et al. 2007), mais aussi dans les démences vasculaires (Roman et al. 2004). L'altération des fonctions exécutives constitue donc une caractéristique neuropsychologique des principales causes de démences : MA, démences associées à la maladie de Parkinson et syndromes parkinsoniens et démence vasculaire. L'oculomotricité et en particulier les mouvements oculaires rapides dirigés pour « atteindre » visuellement (maintenir l'image sur la rétine fovéale) un objet d'intérêt dans l'espace visuel - ou saccades volontaires - sont à l'interface des systèmes moteurs, sous-corticaux frontaux et fronto-pariétaux des processus d'attention, de prise de décision et des processus de mémoire de travail spatiale (Figure 1, Grosbras et al. 2004).

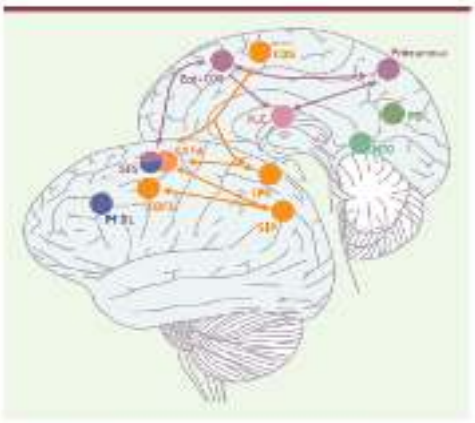


Figure 1 : Grosbras H, Lobel E, Berthoz A M/S 2004 ; 20 :225-30

L'étude des saccades volontaires bénéficie d'un regain d'intérêt dans les pathologies neurodégénératives, en raison :

1) de la simplicité d'enregistrement non invasive grâce aux progrès techniques (video-tracking) et d'analyse automatique du signal,

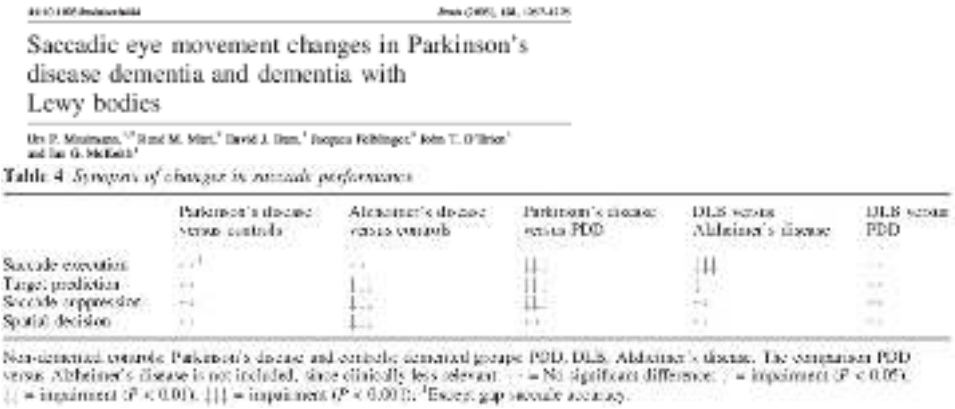
2) de l'absence d'influence de l'état moteur des membres,

3) de l'influence limitée ou connue du vieillissement normal

et 4) de la possibilité de solliciter des ressources cognitives dans des taches de prédiction, d'inhibition, de décision spatiale (ex : anti-saccades, GO/ NO-GO) ou de stratégies d'exploration d'objets ou de scènes.

Les saccades volontaires sont ainsi très précocement altérées dans les pathologies où il existe une altération des fonctions exécutives comme dans la maladie de Parkinson et les affections des noyaux gris centraux, d'autant plus qu'il existe une démence associée (Mosimann et al. 2005, Barker and Michell 2009). Dans la MA, la littérature est moins abondante et plus récente. Les paramètres d'exécution des saccades (latence, vélocité, amplitude) sont peu ou variablement altérés par rapport au processus normal du vieillissement (Abel et al. 2002 ; Boxer et al. 2006 ; Crawford et al. 2005 ; Garbutt et al. 2008 ; Shafiq-Antonacci et al. 2003). Le vieillissement altère de manière assez caractéristique, mais non spécifique, l'amplitude des saccades en particulier verticales, processus qui semble plus marqué dans la MA (Garbutt et al ; 2008) et qui est possiblement lié à une atteinte très précoce du noyau rostral interstitiel du faisceau longitudinal median-riMLF-(Rüb et al 2001). Plus caractéristiques sont les anomalies de suppression des saccades avec non correction des erreurs ou d'échecs dans les taches de prédiction et de décision spatiale (ex : Figure 2) (Crawford et al. 2005 ; Garbutt et al ; 2008 ; Mosimann et al. 2004, 2005). Cependant, seuls des patients avec MA avérée ont été étudiés jusqu'ici. De plus, chez un groupe limité de patients ayant une MA débutante, des anomalies de stratégies de recherche visuelle ont pu être aussi mises en évidence (Rösler et al. 2005). Ces anomalies font intervenir en partie une diminution de la « curiosité » dans l'exploration visuelle, allant de paire avec l'apathie, la passivité et l'indifférence, observées chez ces patients, troubles typiquement dysexécutifs (Daffner et al. 1992).

Figure 2



Saccadic eye movement changes in Parkinson's disease dementia and dementia with Lewy bodies

Table 4 *Synopsis of changes in saccade performance*

	Parkinson's disease versus controls	Alzheimer's disease versus controls	Parkinson's disease versus PDD	DLB versus Alzheimer's disease	DLB versus PDD
Saccade execution	↓↓[1]	↓↓	↓↓	↓↓	↓↓
Target prediction	↓↓	↓↓	↓↓	↓	↓↓
Saccade suppression	↓↓	↓↓	↓↓	↓↓	↓↓
Spatial decision	↓↓	↓↓	↓↓	↓↓	↓↓

Non-demented controls, Parkinson's disease and controls; demented groups: PDD, DLB, Alzheimer's disease. The comparison PDD versus Alzheimer's disease is not included, since clinically less relevant. – = No significant difference; ↓ = impairment ($P < 0.05$); ↓↓ = impairment ($P < 0.01$); ↓↓↓ = impairment ($P < 0.001$); [1]Excess gap saccade accuracy.

Ainsi,

A-L'exécution des saccades volontaires implique des structures très précocement atteintes par le processus pathologique lié à la MA : dégénérescence neurofibrillaire et dépôts de protéine amyloïde.

B- L'altération des saccades volontaires horizontales générées ou inhibées dans des taches impliquant des processus de décision spatiale, de prédiction ou d'exploration visuelle et de curiosité :

- 1) Sont un reflet des troubles dysexécutifs précocement mis en évidence par la neuropsychologie dans la MA (Amieva et al. 2008),
- 2) Impliquent des structures sous-cortico-fronto-pariétales précocement atteintes par le processus pathologique de MA (dépôts amyloïdes) comme le noyau caudé et le pré-cuneus (Brown et al. 2007 ; Scheinin et al. 2009),
- 3) Contribuent aux processus d'attention et d'exploration visuospatiale dont l'altération participe à la survenue de la dégradation des activités de la vie quotidienne dans la MA (Rösler et al. 2000 et 2005, Mosimann et al. 2004).

Nous avons conscience que la définition proposée pour la MAP est en cours de validation par ailleurs. Néanmoins, les entités proposées par Dubois et al. 2010 permettent d'identifier des personnes à un stade précoce et sont utiles pour définir différents stades permettant l'évaluation de notre hypothèse de recherche, très physiopathologique.

1.2. HYPOTHÈSES DE LA RECHERCHE ET RÉSULTATS ATTENDUS

Démontrer par une étude pilote transversale de type « cas-témoins » physiopathologique et diagnostique qu'il existe des anomalies d'exécution des saccades détectables dans les MAP. En nous basant sur les données de la littérature, nous pouvons attendre des anomalies intermédiaires entre des sujets contrôles appariés pour l'âge et des MA typiques avérées dans des tâches visuelles d'anti-saccades, de décision spatiale et de prédiction (Monsimann et al 2005). Ces anomalies pourraient constituer un marqueur clinique dans la sphère motrice au stade prodromal de la maladie. Elles seraient corrélées aux performances neuropsychologiques en particulier dans la composante dysexécutive et pourraient expliquer, avec la dégradation des processus d'exploration visuelle, les anomalies précoces détectées au test des codes (DSST, codes de Wechsler). Ces anomalies peuvent contribuer à l'altération des activités de la vie quotidienne et à la conversion à la démence.

1.3. RAPPORT BÉNÉFICE / RISQUE

Le bénéfice individuel pour les participants est de profiter d'un bilan ophtalmologique avec dépistage des anomalies de l'acuité visuelle, glaucome et pathologies rétiniennes du sujet âgé, mais aussi de bénéficier d'une détection précoce multimodale de MA et de sa prise en charge précoce.

Le bénéfice collectif est la détection précoce multimodale de la MA et sa prise en charge précoce, de contributer à la découverte de biomarqueurs précoces, découverte de nouveaux marqueurs de l'effet thérapeutique de médicaments agissant sur l'évolution initiale de la maladie.

Les risques de cette étude sont inexistants pour les participants : les explorations sont non invasives, et l'imagerie rétinienne est réalisée sans dilatation.

En résumé, le rapport risque/bénéfice est très favorable pour les participants.

1.4. RETOMBÉES ATTENDUES

Les retombées attendues sont :

- La contribution à la détection multimodale de la MA prodromale,
- Biomarqueur précoce de la MA,
- La contribution à la compréhension physiopathologique de l'altération précoce de tests dysexécutifs comme le DSST,
- La contribution à la compréhension de la physiopathologie de l'altération des activités de la vie quotidienne et du risque accidentel (conduite automobile, voie publique, chutes) dans les phases précoces de la MA,
- La définition de nouvelles variables dans l'essai thérapeutique symptomatique ou modifiant le cours évolutif des phases précoces de la MA (Andrieu et al. 2009).

OBJECTIFS DE LA RECHERCHE

1.5. OBJECTIF PRINCIPAL

Comparer la distribution des paramètres d'exécution des saccades (latence, vitesse, précision, erreurs) dans des taches de pro-saccades horizontales et verticales, d'inhibition (anti-saccades), de prédiction et de décision spatiale (selon Monsimann et al. 2005) dans la MAP comparée à des MA probables légères à modérées et à des sujets contrôles appariés (± 5ans) concernant l'âge au groupe MAP.

1.6. OBJECTIFS SECONDAIRES

- Mesurer les altérations des stratégies d'exploration visuelle (test de détection d'items-Rösler et al. 2005 et de curiosité dans l'exploration d'images dégradées de visages ou images incongrues, selon Daffner et al. 1992 et Laughland et al. 2002).

- Etudier l'association entre l'altération des saccades ou les stratégies d'exploration visuelle et curiosité et les performances cognitives en particulier celles obtenues au test des codes de la WAIS (DSST).

- Comparer le nombre moyen de points de fixation dans les régions dégradées et des cartes d'attentions visuelles induites.

CONCEPTION DE LA RECHERCHE

1.7. SCHÉMA DE LA RECHERCHE

Etude nationale multicentrique transversale de type « cas-témoin » de nature physiopathologique comparant 3 groupes de patients :

- Alzheimer prodromal (groupe A),
- Alzheimer typique de forme légère à modérée (Groupe B)
- et sujets contrôles (Groupe C) appariés pour l'âge (± 5ans) au groupe A.

1.8. CRITERES D'ELIGIBILITE

1.8.1. CRITÈRES D'INCLUSION

Tous groupes :

- Age >60 ans, des deux sexes,
- Participants affiliés ou bénéficiaires à un régime de Sécurité Sociale.
- Participants ayant donné leur consentement éclairé. Le consentement peut être recueilli par le représentant du participant, le cas échéant.

Groupe A : Maladie d'Alzheimer prodromale (MAP) ou pré-démentielle (critères Dubois et Albert 2004 et revus par Dubois et al. 2010).

- Plainte mnésique rapportée par le patient ou la famille
- Activités complexes de vie la quotidienne normale (IADL-Lawton) ou légèrement altérées (seulement le premier niveau atteint)
- Syndrome amnésique de type hippocampique défini par un rappel libre très faible malgré un encodage adéquat, diminution du rappel total en raison de l'absence d'effet de l'indiçage (RL/RI-16items [Test de Gröber et Buschke] = rappel libre total inférieur à 17 ou rappel total inférieur à 40) ou trouble de la reconnaissance, intrusions nombreuses)
- Persistance des troubles mnésiques à une évaluation ultérieure (> 3 mois)
- CDR ≤ 0,5
- Absence de démence totalement développée (MMSE ≥ 24) ou suivant l'appréciation du clinicien,
- Exclusion d'autres maladies pouvant causer un trouble cognitif léger (MCI),
- IRM cérébrale montrant une atrophie médio-temporale/hippocampique ou hypo métabolisme temporo-pariétal en TEP ou SPECT ou si LCR disponible (non obligatoire) ratio bétaA42/tau anormal.

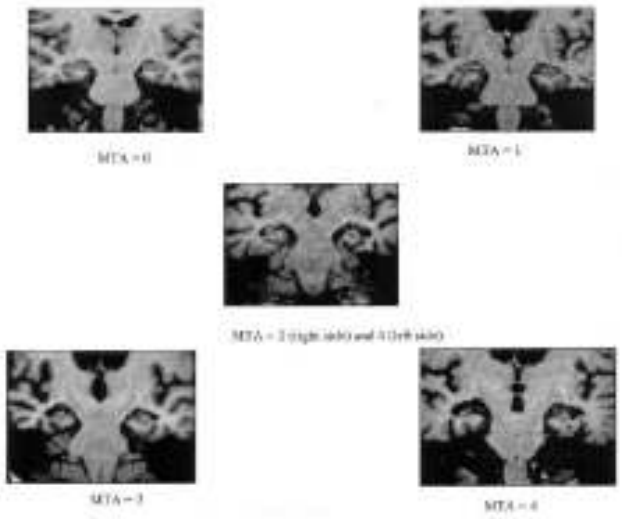


Figure 1. Medial temporal lobe atrophy (MTA) scale. The rating from 0 to 4 is displayed; higher scores indicate more atrophy. When one score is given, left = right.

Table 1 Scheme of medial temporal lobe atrophy rating

Score	Width of choroid fissure	Width of temporal horn	Height of hippocampus
0	Normal	Normal	Normal
1	↑	Normal	Normal
2	↑↑	↑	↓
3	↑↑↑	↑↑	↓↓
4	↑↑↑	↑↑↑	↓↓↓

A score of 0 to 4 is given separately for the left and right side.
(↑) = increase; (↓) = decrease.

Figure 3 : Méthode d'analyse du score MTA de Scheltens et al. 1992, tiré de Korf et al. 2004.

Groupe B : MA typiques légères à modérées

- Critères diagnostiques NINDS-ADRDA actualisés selon Dubois et al. 2007 et 2010.
- MMSE ≥ 20

Groupe C : Sujets contrôles

- Pas de plainte mnésique ou cognitive
- MMSE ≥ 24 si bas niveau d'étude (CEP -)
- MMSE ≥ 26 si haut niveau d'étude (CEP ou +)

1.8.2. CRITÈRES DE NON INCLUSION

Tous groupes :

- participant ne lisant pas P8 de prés avec correction
- Troubles de l'oculomotricité et strabismes à l'appréciation du clinicien
- Dépression (diagnostic clinique ou score ≥ 10 à l'échelle GDS)
- Sujet ne pouvant donner son consentement éclairé

Contrôles :

- Plainte mnésique ou autre plainte cognitive significative.
- Anomalies au bilan neuropsychologique à l'inclusion laissant supposer une dégradation des fonctions cognitives

1.9. MODALITÉS DE RECRUTEMENT

Les sujets ayant déjà reçu le diagnostic de MAP selon les critères définis (Dubois et al. 2010) seront recrutés parmi les patients ayant consulté pour plainte mnésique et ayant eu un bilan neuropsychologique, biologique et d'imagerie IRM diagnostique au Centre de Mémoire (CMRR) et à la consultation Alzheimer des CHU de Bordeaux, de Lyon et de l'hôpital de la Timone à Marseille. Le bilan neuropsychologique ainsi que l'IRM diagnostique de l'étude correspondent au bilan diagnostique habituellement entrepris dans les 3 centres.

La réalisation d'une ponction lombaire à titre diagnostique pour déterminer le rapport bétaA42/tau n'est pas exigée par le protocole. Les sujets ayant reçu le diagnostic de maladie d'Alzheimer typique selon les critères les plus récents (Dubois et al. 2010) seront recrutés dans les consultations spécialisées des 3 centres. Le transport nécessaire pour se rendre au oculomètrique sera pris en charge à hauteur de 90€/ patient.

Un appariement individuel des témoins au groupe MAP selon l'âge +/- 5 ans sera réalisé. Leur motivation sera accentuée par une rémunération forfaitaire de 150 euros comprenant le transport.

La durée d'inclusion sera de 2 ans.

PROCEDURE(S) ASSOCIEES

1.10. PROCÉDURE DE L'ÉTUDE

1.10.1. BILAN NEUROPSYCHOLOGIQUE

Il est réalisé lors de l'étape diagnostique en amont de la recherche. Le bilan choisi correspond à celui habituellement réalisé dans les CMRR dans le cadre du diagnostic et des bonnes pratiques pour le diagnostic de MA. Il ne sera pas refait dans le cadre de la recherche dans le cas où il daterait de moins de 3 mois.

Le bilan neuropsychologique comprend :

- le **Mini Mental Status Examination** (MMSE) permettant d'effectuer une évaluation cognitive globale de manière standardisée (version GRECO).
- des tests appréciant notamment la mémoire verbale épisodique avec un apprentissage, comportant un contrôle de l'encodage, des rappels libres, indicés, immédiats et différés, (RL/RI-16 items [Test de Gröber et Buschke]) (Van der Linden et al. 2003).
- des tests de **reconnaissance visuelle** développés pour le diagnostic précoce de la MA **(DMS 48)** (Barbeau et al 2004)
- des **tests de fluences verbales** (littérale et catégorielle) pour évaluer la mémoire sémantique, de même qu'un type de fonctions exécutives, à savoir l'accès au stock sémantique, les stratégies de recherche active en mémoire, la sélection des items cibles et l'inhibition des items étrangers, et le contrôle attentionnel
- le **Trail Making Test A et B** sollicite la rapidité de la perception, de même que le contrôle exécutif et la mémoire de travail (TMT, Reitan et al, 1956).
- le **Test des codes** (Digit symbol test ou Code de la Wechsler Adult Intelligence Scale (W.A.I.S) permet d'évaluer en quelques minutes la mémoire de travail, la rapidité et la précision grapho-motrice, la gestion de l'espace visuel, la mise en place de stratégies efficientes, etc. (Singh et al, 1997),
- l'échelle **Clinical Dementia Rating Scale** (CDR, Hughes et al, 1982) évalue six domaines explorant deux dimensions : la dimension cognitive (mémoire, orientation, jugement) et la dimension fonctionnelle avec les activités de la vie quotidienne (autonomie dans les activités sociales, loisirs et

activités domestiques, soins personnels). La CDR est sensible pour différencier les patients non déments des déments.

- la **DO 80** est une épreuve de dénomination orale d'images permettant d'évaluer l'aspect de production du langage (Deloche & Hannequin, 1997).
- Le **test des similitudes** est un sous-test de l'échelle d'intelligence de Wechsler (WAIS) permettant d'évaluer les capacités d'abstraction verbale d'un individu (le sujet doit déterminer ce qu'il y a de commun entre 2 items). Ce test est également sensible aux dysfonctionnements de la mémoire sémantique (Singh *et al*, 1997).
- Le **digit span** ou empan de chiffres est un sous-test de la WAIS permettant d'estimer le fonctionnement de la mémoire de travail. Cette épreuve donne lieu au calcul de 2 indices : l'empan de chiffres en ordre direct et l'empan de chiffres en ordre indirect, permettant pour l'un d'évaluer les capacités de stockage passif d'informations en mémoire de travail et pour l'autre la capacité de stockage et de manipulation d'informations en mémoire de travail. (Singh *et al*, 1997).

La batterie de tests sera complétée, pour tous les patients, par un questionnaire d'évaluation des activités instrumentales de vie quotidienne [ADL, IADL (Lawton et Brody, 1969)] et par des échelles d'évaluation de l'anxiété et de la dépression [GDS, Yesavage *et al*, 1982]

1.10.2. BILAN OCULOMÉTRIQUE

Les enregistrements des mouvements oculaires seront réalisés à l'aide du dispositif automatisé non invasif et mobile proposé par la société e(ye)BRAIN, le mobile e(ye)BRAIN TRACKER http://eyebrain.com/ebrainv5/index.php?option=com_content&task=view&id=139&nav=1&lang=en&mid=45&site=tr&smid=3&whr=offr&lang=en&cmid=86.

Ce vidéo-oculomètre à caméra infrarouge permet d'acquérir les mouvements des yeux à une fréquence de 500 Hz en mode binoculaire (enregistrement des deux yeux simultanément) avec une précision spatiale de 0,5°. L'écran est positionné à 60 cm du sujet sur un écran d'au minimum 21 pouces. Les stimulations visuelles sont réalisées avec le logiciel meyeParadigm et les analyses avec le logiciel meyeAnalysis (Figure 4).

Figure 4. Mobile EyeBrainTracker

Les tests oculomoteurs qui seront utilisés sont les pro-saccades (gap) horizontales et verticales (Figures 5 et 6), le test d'anti-saccades (Figure 8), la prédiction, la décision spatiale selon Monsimann et al. 2005 (Figure 9), la recherche d'items selon Rösler et al. 2005 (Figure 10) et des tâches d'exploration et de curiosité d'images et de visages normaux ou incongrus adaptés de Daffner et al. 1992 et Loughland et al. 2002 (Figures 11 et 12)

Les paradigmes utilisés :

- **Gap**

Le but est de tester les saccades réflexes horizontales et verticales
Au cours de ce test, le participant réalise des saccades horizontales visuellement guidées. Ce test est constitué de 12 séquences présentées aléatoirement dont 6 correspondent à des saccades à gauche, 6 à des saccades à droite (Figure 4).

Chaque séquence se décompose en 3 écrans successifs :
Une cible fixe au centre de l'écran pendant une durée variable de 2400 à 3600 ms
Un écran noir (constituant le Gap) pendant 200 ms
Une cible fixe présentée à 22° à gauche ou à droite pendant 1000 ms

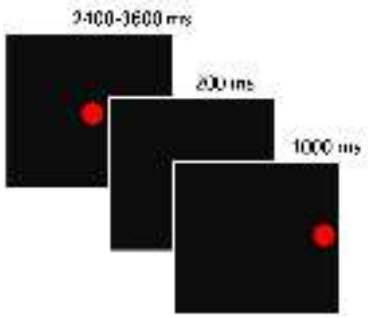


Figure 5 : *Schéma gap*

A la fin du test, une cible centrale est présentée pour conclure.

Ce test est présenté 2 fois au sujet avec des séquences aléatoires différentes. Le test dure une minute.

- **Saccades verticales**

Au cours de ce test, le sujet réalise des saccades verticales visuellement guidées. Ce test est constitué de 12 séquences présentées aléatoirement dont 6 correspondent à des saccades vers le haut, 6 à des saccades vers le bas (Figure 6).
Chaque séquence se décompose en 2 écrans successifs :
Une cible fixe au centre de l'écran pendant une durée variable de 2400 à 3600 ms
Une cible fixe présentée à 13° vers le haut ou vers le bas pendant 1000 ms

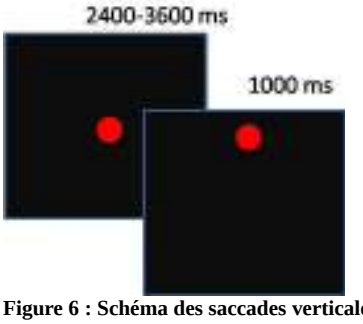


Figure 6 : Schéma des saccades verticales

A la fin du test, une cible centrale est présentée pour conclure.

Ce test est présenté 2 fois au sujet avec des séquences aléatoires différentes. Le test dure une minute.

Les variables de base analysées sont (Figure 7) :
Les paramètres des saccades :

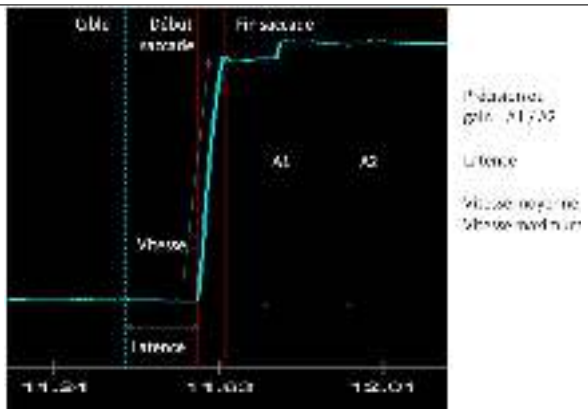


Figure 7 : enregistrement d'une saccade

Le temps de latence (msec) ou temps entre l'apparition de la cible et le début de la saccade vers la cible vers la droite et vers la gauche.
La vitesse moyenne à droite et à gauche ou en haut et en bas (degré/sec). La vitesse moyenne est la vitesse moyenne des saccades entre le début et la fin de la saccade en degré par secondes.
Le gain. Le calcul du gain est le rapport entre la ligne de base de et le premier plateau après la saccade (A1) et la ligne de base et le second plateau (A2).

- **Anti-saccades**

La tâche consiste à tester la façon dont le sujet inhibe le regard en direction d'une « non-cible » et dirige son regard du côté opposé.
Le test anti-saccades comprend des saccades horizontales à réaliser dans la direction opposée de la cible.

Ex : une cible est présentée à gauche, le participant doit regarder directement à droite (Figure 8).
Ce test est constitué de 12 séquences présentées aléatoirement dont 6 correspondent à des anti-saccades à gauche, 6 à des anti-saccades à droite.
Chaque séquence se décompose en 3 écrans successifs :
Une cible fixe au centre de l'écran pendant une durée variable de 2400 à 3600 ms
Un écran noir pendant 200 ms
Une cible fixe présentée à 22° à gauche ou à droite pendant 1000 ms

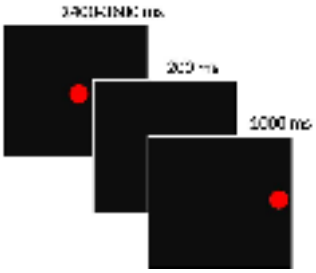


Figure 8 Schéma des anti-saccades

A la fin du test, une cible centrale est présentée pour conclure.

Ce test est présenté 2 fois au sujet avec des séquences aléatoires différentes. Le test dure une minute.

Les variables analysées sont :
La latence de déclenchement (msec)

Le pourcentage d'erreurs (mauvaise direction)
Le pourcentage d'erreurs corrigées

Les variables utilisées pour l'analyse seront celles de l'œil le moins artéfacté et dans le cas où les deux yeux auront un bon tracé, nous prendrons en compte celui qui aura la meilleur latence de déclenchement.

- **Prédiction (Mosimann 2005) Figure 9.**

Dans cette tâche la direction de la cible (droite, centre, gauche), son angulation (16°) et la durée du stimulus (1000 msec) sont entièrement prévisibles. Le but est de calculer combien de fois les sujets sont capable de prédire la position suivante de la cible. Comme plus de 80 msec sont nécessaires pour percevoir le stimulus visuel, toute réponse < 80 msec est considéré comme prédite.

Variables analysées :
Le pourcentage de prédiction
La latence
La précision

- **Décision (Mosimann 2005) Figure 9.**

Le test analyse la capacité du sujet à faire un jugement spatial et prendre une décision. Deux cibles vertes sont présentées simultanément à différents angles de chaque côté de la position centrale aléatoirement à droite et à gauche par paire (12 et 16°, 16 et 8°, 12 et 8°). Le stimulus central est présenté 1700-2900 msec puis les deux cibles 1500 msec. Le sujet doit regarder la cible la plus près du point central, la cible la plus éloignée du centre devient rouge indiquant l'erreur.

Les variables analysées sont :
Le pourcentage d'erreurs
La latence moyenne des saccades
Le pourcentage d'erreurs corrigées

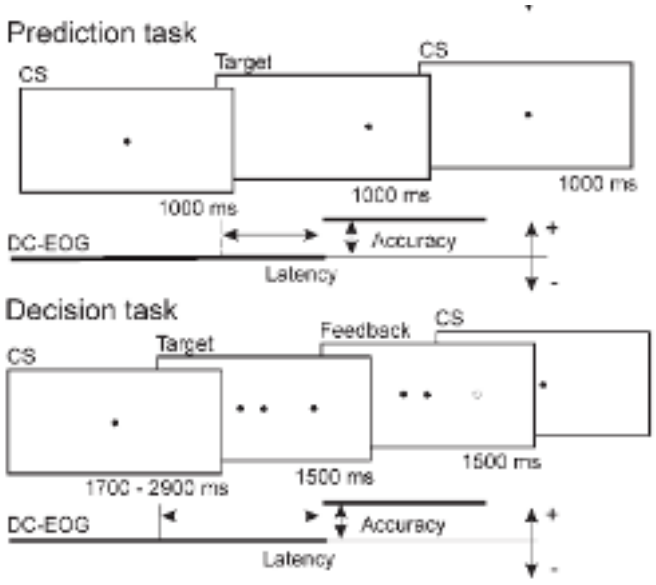


Figure 9 : Paradigme de décision spatiale et de prédiction selon Monsimann et al 2005

- **Tâches d'exploration visuelle**

Détection d'items, tâche de recherche visuelle (Figure 10).
Le paradigme employé dérive de Rösler et al. 2005. En bref, une croix de fixation est proposée au sujet puis un objet cible pendant 1000 msec puis une présentation de 6 objets. Le sujet doit presser un bouton dès l'objet vu et citer l'objet.

Les variables sont :

La latence moyenne de la détection de la cible

Le nombre de fixations erronées et la distribution des fixations (centrale ou périphérique à la région d'intérêt).

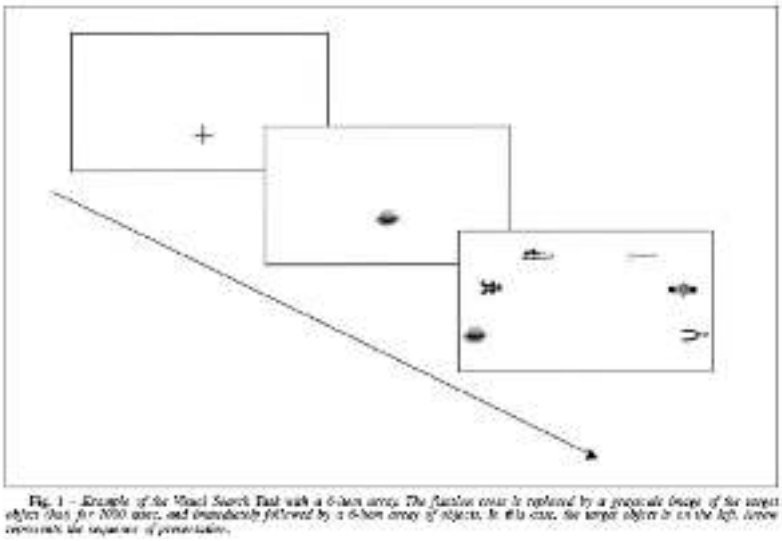


Figure 10 : Tâche de reconnaissance d'items selon Rösler et al. 2005

- **Curiosité/stratégies d'exploration**

Dans un paradigme adapté de Daffner et al. 1992, des paires d'images dont l'une est incongrue sont présentées au sujet. Des tâches similaires sont effectuées avec l'exploration de visages normaux ou dégradés selon Loughland et al. 2002. La variable enregistrée est la durée de fixation sur les images qui est chez le sujet normal supérieur sur les images incongrues que sur les images normales, alors que les sujets atteints de MA passent moins de temps sur les images incongrues, nouvelles et curieuses.

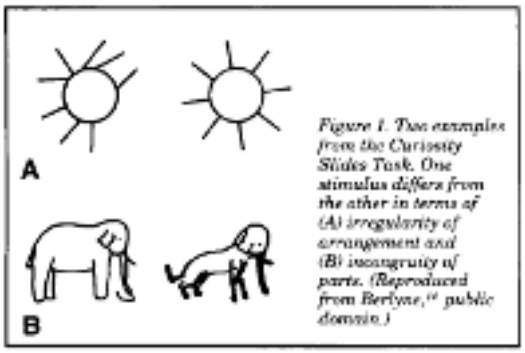


Figure 11 : Tâche d'exploration d'images et d'images incongrues selon Daffner et al. 1992

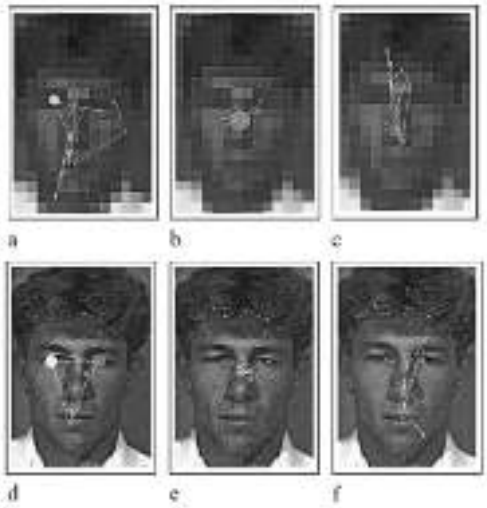


Figure 12 : Tâche d'exploration de visages ou d'images de visages dégradés selon Loughland et al. 2002

Dans un second temps, des vidéos naturelles floutées et non dégradées sont projetés au sujet. Cette tâche consiste à projeter des séquences de 28 secondes séparée d'un écran noir de 200 msec. L'examen est réalisé en deux temps ; un premiers temps deux séries de vidéos « normales » non dégradées sont présentées au sujet puis dans un second temps deux séries de vidéos « floutées ».

CRITERES DE JUGEMENT

1.11. CRITÈRE DE JUGEMENT PRINCIPAL

- Paramètres d'exécution des saccades :
- La latence moyenne de déclenchement (msec).
- Vitesse moyenne (°/msec) et vitesse maximale.
- Précision ou gain moyen (ratio composante A1/A2).
- Le pourcentage moyen d'erreurs et pourcentages d'erreurs corrigées.
- Pourcentage moyen de prédiction.

1.12. CRITÈRES DE JUGEMENT SECONDAIRES

- Scores aux différents tests du bilan neuropsychologique.
- Variables définies lors des tests de détection d'items et d'exploration des visages et d'images incongrues (nombre de fixations, durée des fixations, erreurs).
- Nombre moyen de points de fixations dans les régions dégradées et des cartes d'attentions visuelles induites

DEROULEMENT DE LA RECHERCHE

1.13. CALENDRIER DE LA RECHERCHE

- Durée de la période d'inclusion : 24 mois
- Durée de participation de chaque participant : 1 mois maximum
- Durée totale de la recherche : 30 mois dont 3 premiers mois de phase préparatoire, 24 mois d'étude et 3 mois d'analyse des données et rédaction du rapport final.

1.14. TABLEAU RÉCAPITULATIF DU SUIVI PARTICIPANT

	Inclusion V0	Visite d'étude V1
Consentement éclairé	✓	
Critères d'inclusion/non inclusion	✓	
Bilan neuropsychologique[1]	✓	
MMSE et ADL	✓	✓
Oculométrie		✓

[1]si date de plus de 3 mois

La visite d'inclusion est assurée par le médecin investigateur. Le participant sera informé en amont de la nature de l'étude et recevra une notice d'information.
La visite d'inclusion a lieu entre 0 et 1 mois avant la visite d'étude.
Avant tout examen lié à la recherche, l'investigateur recueille le consentement libre, éclairé et écrit du participant (ou de son représentant légal le cas échéant).
Le bilan neuropsychologique ne sera réalisé chez les sujets des groupes A et B que dans le cas où il date de plus de 3 mois, il sera réalisé systématiquement chez les sujets contrôles. Un MMSE et les activités de la vie quotidienne seront contrôlés avant l'examen oculométrique pour vérifier l'absence de conversion très improbable des sujets atteints de MAP durant la période maximale de 3 mois après le dernier examen neuropsychologique.

RECUEIL DU CONSENTEMENT

Lors de la visite d'inclusion, le médecin investigateur informe le participant et répond à toutes ses questions concernant l'objectif, la nature des contraintes, les risques prévisibles et les bénéfices attendus de la recherche. Il précise également les droits du participant dans le cadre d'une recherche biomédicale et vérifie les critères d'éligibilité. Un exemplaire de la note d'information et du formulaire de consentement est alors remis au participant par le médecin investigateur.

Après cette séance d'information, le participant dispose d'un délai de réflexion. Le médecin investigateur est responsable de l'obtention du consentement éclairé écrit du participant. Le formulaire de consentement doit être signé AVANT LA REALISATION DE TOUT EXAMEN CLINIQUE OU PARACLINIQUE NECESSITE PAR LA RECHERCHE.

Si le participant donne son accord de participation, ce dernier et l'investigateur inscrivent leurs noms et prénoms en clair, datent et signent le formulaire de consentement.

En accord avec les critères d'inclusion et de non inclusion, les participants éligibles ne présenteront pas de dégradation des fonctions cognitives suffisamment sévères pour altérer leur capacité à consentir de manière informée. Cependant, par prudence, l'accord d'une personne représentante pourra être sollicité dans le groupe MA. Les différents exemplaires de la note d'information et du formulaire de consentement sont alors repartis comme suit :

- Un exemplaire de la note d'information et du consentement signé sont remis au participant.
- L'exemplaire original est conservé par le médecin investigateur (même en cas de déménagement du participant pendant la durée de la recherche) dans un lieu sûr inaccessible à des tiers.
- A la fin des inclusions ou au plus tard à la fin de la recherche, un exemplaire de chaque formulaire de consentement est transmis au promoteur ou à son représentant selon des modalités communiquées en temps utile aux investigateurs.

VISITE D'INCLUSION (V0) :

Durée 2 heures (si bilan neuropsychologique)

- Consentement éclairé
- Critères d'inclusion et de non-inclusion
- Bilan neuropsychologique (CMRR) : passation des tests suivants : MMSE (version Greco), RL/RI-16items (Van der Linden 2003), Test de reconnaissance visuelle (DMS48), fluence verbale littérale et catégorielle, TMT A et B (Reitan 1958), Test des codes (DSST, Weschler 1997), Clinical Dementia Rating Scale (Hughes 1982), dénomination d'images DO80 (Deloche et Hannequin 1997), Test des similitudes et Digit Span de la WAIS, symptômes dépressifs (GDS), si date de plus de 3 mois (bilan effectué à titre diagnostique) chez les sujets des groupes A et B et systématiquement chez les sujets contrôles.

VISITE DE SUIVI (V1)

Durée 1 heure et 45 minutes, dans le mois suivant la V0.

- Contrôle du MMSE et des ADL
- Oculométrie automatisée non-invasive 45 minutes avec pauses : poursuite, saccades horizontales, saccades verticales, anti-saccades, prédiction, décision spatiale (Monsiman et al. 2005), détection *d'Items* (Rösler et al. 2005) et exploration de visages, images incongrues ainsi que de vidéos naturelles dégradées et non dégradées projetées selon Daffner et al. 1992 et Loughland et al 2002.

1.15. ABANDON ET RETRAIT DE CONSENTEMENT

Le participant qui souhaite abandonner ou retirer son consentement de participation à la recherche (comme il est en droit de le faire à tout moment), n'est plus suivi dans le cadre du protocole, mais doit faire l'objet de la meilleure prise en charge possible compte tenu de son état de santé et de l'état des connaissances du moment.

Un **abandon** est une décision d'un participant inclus de faire valoir son droit d'interrompre sa participation à une recherche, à tout moment au cours du suivi, sans qu'elle n'encoure aucun préjudice de ce fait et sans avoir à se justifier.

L'investigateur doit identifier la cause de l'abandon et évalue s'il est possible de recueillir la variable sur laquelle porte le critère de jugement principal au moment de l'abandon. Les abandons de recherche doivent être notifiés rapidement au centre investigateur coordonnateur par fax. Les raisons et la date d'abandon doivent être documentées dans le cahier d'observation.

Un **retrait de consentement** est une décision d'un participant de revenir sur sa décision de participer à une recherche et de faire valoir son droit d'annuler son consentement éclairé, à tout moment au cours du suivi et sans qu'il n'encoure aucun préjudice de ce fait et sans avoir à se justifier.

Lorsqu'un participant retire son consentement de participation à la recherche, l'investigateur doit contacter le centre investigateur coordonnateur et le centre de méthodologie et de gestion des données. Les données concernant le participant sont retirées de la base de données conformément à la loi relative à l'informatique, aux fichiers et aux libertés et les échantillons biologiques sont détruits.

1.16. RÈGLES D'ARRÊT DE LA RECHERCHE

Fin de la recherche ou arrêt prévu de la recherche : terme de la participation de la dernière personne qui se prête à la recherche (cf. Articles L.1123-11 ; R.1123-59 du Code de Santé Publique) aussi appelé dernière visite du dernier participant inclus dans la recherche.

Cette définition est proposée par défaut dans le cadre de la Loi de Santé Publique. Toute autre définition doit être mentionnée dans le protocole.

Lorsque la recherche a atteint son terme prévu (arrêt prévu), la fin de la recherche doit être déclarée à l'ANSM dans un délai de 90 jours.

Arrêt anticipé de la recherche : la recherche clinique est arrêté (définitivement) de façon anticipée. C'est le cas, notamment, lorsque le promoteur décide :

- de ne pas commencer la recherche malgré l'obtention de l'autorisation de l'ANSM et de l'avis favorable d'un CPP ;
- de ne pas reprendre la recherche après l'avoir interrompu temporairement ou après sa suspension par l'ANSM.

Lorsque la recherche est arrêtée (définitivement) de façon anticipée, la fin de la recherche doit être déclarée à l'ANSM dans un délai de 15 jours en indiquant les raisons qui motivent cet arrêt.

Arrêt temporaire de la recherche (cf. Article R.1123-55 du CSP ; Arrêté MS-HPS (article 4)) : l'arrêt temporaire d'une recherche clinique consiste en :

- l'arrêt de l'inclusion de nouvelles personnes dans cette recherche;
- et/ou l'arrêt de l'administration du produit testé, le cas échéant, à tout ou partie des personnes déjà incluses dans la recherche ;
- et/ou l'arrêt de la pratique des actes prévus par le protocole de la recherche.

Toute décision du promoteur d'interrompre temporairement la recherche doit faire l'objet d'une information immédiate à l'ANSM et au CPP concerné et dans un second temps et dans un délai maximum de 15 jours calendaires suivant la date de cette interruption, d'une demande d'autorisation de modification substantielle concernant cet arrêt temporaire soumise à l'ANSM et d'une demande d'avis au CPP concerné.

1.17. DÉVIATIONS AU PROTOCOLE

Les déviations peuvent concerner tous les aspects d'un protocole de recherche : processus d'inclusion, suivi, mesure des critères de jugement, traitements. Toutes doivent être documentées par l'investigateur et discutées en Conseil Scientifique.

Seuls les abandons entrainent un arrêt du suivi. Même en cas de déviation au protocole, le suivi du participant doit être mené jusqu'au terme prévu dans le protocole.

1.17.1. PATIENT PERDU DE VUE

Un participant est considéré comme perdu de vue quand il arrête le suivi prévu dans le cadre du protocole sans raison connue de l'investigateur, de sorte que le recueil des données ne peut pas être effectué comme prévu.
Les participants perdus de vue doivent faire l'objet d'une recherche active de la part de l'investigateur.

1.17.2. PARTICIPANTS INCLUS À TORT

Un participant est considéré comme inclus à tort lorsqu'il a effectivement été inclus dans la recherche alors qu'il ne vérifiait pas tous les critères d'éligibilité. Les participants inclus à tort doivent faire l'objet d'une discussion en Conseil Scientifique. Ils doivent continuer à être suivis comme prévu par le protocole jusqu'à ce qu'une décision ait été prise par le Conseil Scientifique.

CONTRAINTES LIÉES À LA RECHERCHE ET INDEMNISATION ÉVENTUELLE DES PARTICIPANTS.

- Réalisation d'un bilan neuropsychologique d'une durée de 2 heures avec des pauses si ce bilan date de moins de 3 mois chez les sujets malades (MAP et MA) et systématique chez les sujets contrôles.
- Réalisation d'un bilan oculomètrique non invasif durant 1h au maximum avec pauses.
- Les explorations réalisées sont non invasives et sans danger pour la santé.
- Les sujets contrôle sont indemnisés à hauteur de 150 euros (participation à la recherche et les trajets), ceux-ci seront inscrit dans le fichier national des personnes qui se prêtent à des recherches biomédicales.

GESTION DES ÉVÉNEMENTS INDÉSIRABLES ET DES FAITS NOUVEAUX

DÉFINITIONS

Evénement indésirable (article R.1123-39 du code de la santé publique)
Toute manifestation nocive survenant chez une personne qui se prête à une recherche biomédicale, que cette manifestation soit liée ou non à la recherche ou au produit sur lequel porte cette recherche.

Evénement indésirable grave (article R.1123-39 du code de la santé publique et guide ICH E2B)
Tout événement indésirable qui :

- ✓ Entraîne la mort,
- ✓ Met en danger la vie de la personne qui se prête à la recherche,
- ✓ Nécessite une hospitalisation ou la prolongation de l'hospitalisation,
- ✓ Provoque une incapacité ou un handicap important(e) ou durable,
- ✓ Se traduit par une anomalie ou une malformation congénitale,
- ✓ Ou tout événement considéré médicalement grave,

et s'agissant du médicament, quelle que soit la dose administrée.

L'expression « mettre en danger la vie » est réservée à une menace vitale immédiate, au moment de l'événement indésirable, et ce, indépendamment des conséquences qu'aurait une thérapie correctrice ou palliative.

Effet indésirable inattendu (article R.1123-39 du code de la santé publique)
Tout effet indésirable dont la nature, la sévérité ou l'évolution ne concorde pas avec les informations relatives aux actes pratiqués, et méthodes utilisées au cours de la recherche. L'évaluation du caractère inattendu d'un effet indésirable se fait sur la base des informations décrites dans le protocole, relatives notamment, le cas échéant, aux actes et méthodes pratiqués au cours de la recherche.

Fait nouveau (arrêté du 24 mai 2006)
Nouvelle donnée de sécurité, pouvant conduire à une réévaluation du rapport des bénéfices et des risques de la recherche, ou qui pourrait être suffisant pour envisager des modifications des documents relatifs à la recherche, de la conduite de la recherche.

DESCRIPTION DES ÉVÉNEMENTS INDÉSIRABLES GRAVES ATTENDUS

Les événements indésirables graves attendus sont ceux liés à la pathologie à l'étude à ses complications et aux traitements de celles-ci. Ces événements indésirables graves ne seront pas liés aux procédures de la recherche. Ils ne devront pas être notifiés à l'unité de vigilance mais rapportés dans le cahier d'observation du participant.
Tout autre événement indésirable grave devra être notifié sans délai à l'unité de vigilance (voir paragraphe 7.3.).

1.18. CONDUITE À TENIR EN CAS D'EVENEMENT INDESIRABLE GRAVE OU DE FAIT NOUVEAU

L'investigateur doit notifier au promoteur, sans délai à partir du jour où il en a connaissance, tout événement indésirable grave ou tout fait nouveau (à l'exception de ceux définis au paragraphe 7.2.), s'il survient :

- à partir de la date de signature du consentement,
- pendant toute la durée de suivi du participant prévue par la recherche,
- jusqu'à 24 heures après la visite V1
- après la fin du suivi du participant prévue par la recherche, lorsqu'il est susceptible d'être dû à la recherche.

Type d'événement	Modalités de notification	Délai de notification au promoteur
EI non grave	Dans le cahier d'observation	Pas de notification immédiate
EIG attendu	Formulaire de déclaration d'EIG initiale + rapport écrit si nécessaire	Notification immédiate au promoteur à l'exception de ceux définis au paragraphe 7.2.
EIG inattendu	Formulaire de déclaration d'EIG initiale + rapport écrit si nécessaire	Notification immédiate au promoteur à l'exception de ceux définis au paragraphe7.2
Fait nouveau	Formulaire de déclaration + rapport écrit si nécessaire	Notification immédiate au promoteur
Grossesse	Formulaire de déclaration d'une grossesse	Dès confirmation de la grossesse

Unité de Vigilance des Essais Cliniques du CHU de Bordeaux

Tel : 05.57.82.08.34
Fax : 05.57.82.12.62
Courriel : vigilance.essaiscliniques@chu-bordeaux.fr

L'unité de vigilance détermine l'imputabilité de l'événement à l'étude et la nécessité d'une déclaration aux autorités compétentes.

Tous ces événements devront être suivis jusqu'à la complète résolution. Un complément d'information (fiche de déclaration complémentaire) concernant l'évolution de l'événement, si elle n'est pas mentionnée dans le premier rapport, sera envoyé à l'unité de vigilance par l'investigateur.

1.19. DÉCLARATION ET ENREGISTREMENT DES ÉIG INATTENDUS ET DES FAITS NOUVEAUX

L'unité de vigilance déclare sans délai les Effets indésirables graves inattendus et les faits nouveaux survenus au cours de la recherche :

- à l'ANSM,
- au Comité de Protection des Personnes compétent. Le comité s'assure, si nécessaire, que les sujets participant à la recherche ont été informés des effets indésirables et qu'ils confirment leur consentement.

1.20. RAPPORT ANNUEL DE SECURITE (RAS)

A la date anniversaire de la première inclusion, l'unité de vigilance rédige un rapport de sécurité comprenant :

- la liste des effets indésirables graves susceptibles d'être liés à la recherche incluant les effets graves attendus et inattendus,
- une analyse concise et critique de la sécurité des participants se prêtant à la recherche.

Ce rapport est envoyé à l'ANSM et au CPP dans les 60 jours suivant la date anniversaire de la première inclusion.

ASPECTS STATISTIQUES

1.21. CALCUL DE LA TAILLE D'ÉTUDE

Dans cet essai, il y a 3 groupes de participants :

- Groupe MAP
- Groupe MA
- Groupe contrôle

L'objectif principal de cette étude est de comparer les paramètres d'exécution de saccades dans un premier temps entre le groupe MAP et le groupe contrôle, puis entre le groupe MAP et le groupe MA.
Le pourcentage d'erreur aux anti-saccades étant la variable la plus discriminante des paramètres d'exécution il sera utilisé pour le calcul de la taille d'étude.

On fait l'hypothèse que dans le groupe MAP, le pourcentage d'erreur aux anti-saccades est de 60 ± 30 (moyenne ± sd) et dans le groupe contrôle 25 ± 38. Dans cette situation, le calcul fondé sur la formule du test de Student avec un risque α de 5% et une puissance 1-β de 80%, il faut inclure au moins 20 participants dans chaque groupe.

De plus cet effectif permettra de montrer une différence entre le groupe MA (80 ± 42) et le groupe contrôle (25 ± 38) avec une puissance de 80% et un risque alpha de 5%.

Tenant compte de la littérature, du calcul de la taille d'étude, du nombre de centres et de leur capacité de recrutement, il est décidé d'inclure 30 participants par groupe soit 90 participants au total.

1.22. LOGICIELS UTILISES

Les analyses seront réalisées avec le logiciel SAS® (version n°9.1).

1.23. PLAN D'ANALYSE

1.23.1. DESCRIPTION DES INCLUSIONS ET DU SUIVI

Les participants pré-inclus et non inclus dans la recherche seront décrits et comparés aux participants inclus.

Le nombre de participants inclus et la courbe des inclusions (évolution du nombre des participants inclus entre la première et la dernière inclusion) seront présentés par groupe. La durée cumulée de suivi sera calculée (somme du temps de participation pour chacun des participants inclus, c'est-à-dire de la différence en nombre de jours, entre la date d'inclusion et la date des dernières nouvelles dans la recherche) et le rapport durée cumulée effective de suivi/durée cumulée attendue sera présenté.

1.23.2. PARTICIPANTS INCLUS DANS L'ANALYSE

Ne pourront être exclus de l'analyse que les participants qui présentent au moins une des conditions suivantes :

- Participants inclus à tort pour consentement non signé,
- Participants inclus à tort pour critères d'éligibilité non respectés,
- Participants ayant retiré leur consentement.

Cette décision d'exclusion sera prise par le conseil scientifique après documentation de l'observation par le Centre de Méthodologie et de Gestion des données, en insu de l'évolution du participant après l'inclusion.

En dehors de ces exclusions, les participants décédés, perdus de vue ou ayant abandonné la recherche seront tous inclus dans l'analyse.

1.23.3. CARACTÉRISTIQUES DES PARTICIPANTS

Les participants seront décrits selon les variables suivantes :

- Respect des critères d'éligibilité
- Caractéristiques épidémiologiques
- Caractéristiques cliniques
- Caractéristiques de traitements

Une description des violations du protocole et des participants répartis selon ces violations sera faite.

Une description des causes de décès et d'abandon sera faite et les participants décédés, perdus de vue ou ayant abandonné la recherche seront décrits et comparés aux autres participants.

1.23.4. CRITÈRE DE JUGEMENT PRINCIPAL

L'objectif principal est de comparer la distribution des paramètres d'exécution des saccades, dans un premier temps entre le groupe MAP et le groupe contrôle puis entre le groupe MAP et le groupe MA. Les paramètres étant des variables quantitatives, ils seront décrits en termes d'effectif, moyenne, écart-type et intervalle de confiance à 95% de la moyenne, médiane, minimum, maximum, Q1 et Q3.

Les comparaisons 2 à 2 entre les groupes des différents paramètres utiliseront le test paramétrique de Student pour données appariées (comparaison MAP/contrôle) ou non appariées (comparaison MAP/MA) ou le test non paramétrique de Wilcoxon pour données appariées (comparaison MAP/contrôle) ou non appariées (comparaison MAP/MA) si l'hypothèse de normalité n'est pas vérifiée

1.23.5. CRITERES DE JUGEMENT SECONDAIRES

Les altérations des stratégies d'exploration visuelle sont des variables quantitatives et seront décrites en termes d'effectif, moyenne, écart-type et intervalle de confiance à 95% de la moyenne, médiane, minimum, maximum, Q1 et Q3.
Les comparaisons 2 à 2 entre les groupes des différents paramètres utiliseront le test paramétrique de Student pour données appariées (comparaison MAP/contrôle) ou non appariées (comparaison MAP/MA) ou le test non paramétrique de Wilcoxon pour données appariées (comparaison MAP/contrôle) ou non appariées (comparaison MAP/MA) si l'hypothèse de normalité n'est pas vérifiée.
L'association entre l'altération des saccades ou les stratégies d'exploration visuelle et les performances cognitives en particulier celles obtenues au DSST sera étudiée à l'aide d'un coefficient de corrélation de Pearson ou Spearman si l'hypothèse de normalité n'est pas vérifiée.

Un plan d'analyse détaillé sera défini et fera l'objet d'une validation par le conseil scientifique de l'étude.

SURVEILLANCE DE LA RECHERCHE

1.24. CONSEIL SCIENTIFIQUE

1.24.1. COMPOSITION

Il sera composé des personnes suivantes : Pr F. Tison (président), Pr A Vighetto, Dr C. Tilikete, Pr JF Dartigues, Dr H Amieva, Dr MB Rougier, Pr G. Chène (méthodologiste), J. Asselineau , un représentant du promoteur et de l'unité de vigilance.

1.24.2. RYTHME DES RÉUNIONS

Le Conseil Scientifique de la recherche se réunit avant le démarrage de la recherche puis au moins deux fois par an jusqu'à la clôture de la recherche, dont une fois entre un et deux mois après la date anniversaire de l'avis favorable donné à la recherche par l'autorité de santé pour validation du rapport annuel de sécurité.

1.24.3. RÔLE

- Il a pour mission de prendre toute décision importante à la demande de l'investigateur coordonnateur concernant la bonne marche de la recherche et le respect du protocole.
- Il vérifie le respect de l'éthique.
- Il s'informe auprès du centre investigateur coordonnateur de la recherche de l'état d'avancement de la recherche, des problèmes éventuels et des résultats disponibles.
- Il décide de toute modification pertinente du protocole nécessaire à la poursuite de la recherche, notamment :
    - les mesures permettant de faciliter le recrutement dans la recherche,
    - les amendements au protocole avant leur présentation au CPP et à l'autorité de santé compétente,
    - les décisions d'ouvrir ou de fermer des sites participant à la recherche,

- les mesures qui assurent aux personnes participant à la recherche la meilleure sécurité,
- la discussion des résultats et la stratégie de publication de ces résultats.

- Le Conseil Scientifique peut proposer de prolonger ou d'interrompre la recherche en cas de rythme d'inclusion trop lent, d'un trop grand nombre de perdus de vue, de violations majeures du protocole ou bien pour des raisons médicales et/ou administratives. Il précise les modalités éventuelles du suivi prolongé des participants inclus dans la recherche.
- A l'issue de la réunion, le président du Conseil Scientifique doit informer le promoteur des décisions arrêtées. Les décisions concernant un amendement majeur ou une modification de budget doivent être approuvées par le promoteur.

1.25. CENTRE DE MÉTHODOLOGIE ET DE GESTION DES DONNÉES

Le Centre de Méthodologie et de Gestion des données est l'Unité de Soutien Méthodologique à la Recherche clinique et épidémiologique du CHU de Bordeaux (USMR) (méthodologiste de la recherche : Pr Geneviève Chêne).

L'équipe projet est composée d'un méthodologiste, d'un statisticien, d'un data manager et d'un analyste programmeur.

Le Centre de Méthodologie et de Gestion des données :

- collabore à la conception du protocole avec l'investigateur coordonnateur et supervise la conception méthodologique de la recherche,
- finalise la rédaction du protocole avant soumission au CPP et à l'ANSM
- il supervise la conception du cahier d'observation,
- réalise et gère la base de données informatique dédiée à la recherche,
- prépare en collaboration avec l'ARC de centre investigateur coordonnateur et l'ARC du promoteur la mise en place et le suivi de la recherche,
- effectue l'analyse statistique des données,
- participe aux publications et autres valorisations des résultats de la recherche,

Le Centre de Méthodologie et de Gestion des données, en collaboration avec l'ARC de centre coordonnateur, participe à la préparation des résumés présentant l'état d'avancement de la recherche pour le Conseil Scientifique.

Il informe le Conseil Scientifique du déroulement de la recherche et, en collaboration avec l'ARC du centre coordonnateur et du promoteur, participe à la préparation des réunions du comité scientifique et des investigateurs.

1.26. CENTRE INVESTIGATEUR COORDONNATEUR

Le centre investigateur coordonnateur est situé dans le service de Neurologie, Pr F Tison, GH Sud, CHU de Bordeaux et travaille en collaboration avec le promoteur et le Centre de Méthodologie et de Gestion des données.

Le centre coordonnateur est chargé de l'organisation, de la logistique et du monitorage de la recherche. Il informe mensuellement le promoteur sur le déroulement de la recherche (inclusions et suivis, déviations au protocole, propositions d'amendements,…).

1.27. COMITÉ INDÉPENDANT DE SURVEILLANCE

La recherche ne nécessite pas de surveillance particulière en terme de risque.

Ces procédures n'entraînent pas de risque particulier pour les sujets témoins ni pour les personnes présentant la maladie d'Alzheimer. En conséquence, la constitution d'un comité indépendant de surveillance n'est pas nécessaire.

DROITS D'ACCÈS AUX DONNÉES ET DOCUMENTS SOURCE

1.28. ACCÈS AUX DONNÉES

Le promoteur est chargé d'obtenir l'accord de l'ensemble des parties impliquées dans la recherche afin de garantir l'accès direct à tous les lieux de déroulement de la recherche, aux données source, aux documents source et aux rapports dans un but de contrôle de qualité et d'audit par le promoteur.

Les investigateurs mettront à disposition les documents et données individuelles strictement nécessaires au suivi, au contrôle de qualité et à l'audit de la recherche biomédicale, à la disposition des personnes ayant un accès à ces documents conformément aux dispositions législatives et réglementaires en vigueur (articles L.1121-3 et R.5121-13 du code de la santé publique).

1.29. DONNÉES SOURCE

Tout document ou objet original permettant de prouver l'existence ou l'exactitude d'une donnée ou d'un fait, enregistrés au cours de la recherche est défini comme document source.
Le recueil des tests neuropsychologiques est directement reporté dans le cahier d'observation qui sera considéré comme donnée source.

1.30. CONFIDENTIALITÉ DES DONNÉES

Conformément aux dispositions législatives en vigueur (articles L.1121-3 et R.5121-13 du code de la santé publique), les personnes ayant un accès direct aux données source prendront toutes les précautions nécessaires en vue d'assurer la confidentialité des informations relatives aux recherches, aux personnes qui s'y prêtent et notamment en ce qui concerne leur identité ainsi qu'aux résultats obtenus. Ces personnes, au même titre que les investigateurs eux-mêmes, sont soumises au secret professionnel.

Pendant la recherche biomédicale ou à son issue, les données recueillies sur les personnes qui s'y prêtent et transmises au promoteur par les investigateurs (ou tous autres intervenants spécialisés) seront rendues anonymes. Elles ne doivent en aucun cas faire apparaître en clair les noms des personnes concernées ni leur adresse.

Seule la première lettre du nom et du prénom du sujet seront enregistrées, accompagnées d'un numéro codé propre à la recherche indiquant l'ordre d'inclusion des sujets.

Le promoteur s'assurera que chaque personne qui se prête à la recherche a donné son accord par écrit pour l'accès aux données individuelles la concernant et strictement nécessaires au contrôle de qualité de la recherche.

CONTRÔLE ET ASSURANCE QUALITÉ

1.31. CONSIGNES POUR LE RECUEIL DES DONNÉES

Toutes les informations requises par le protocole doivent être consignées sur les cahiers d'observation et une explication doit être apportée pour chaque donnée manquante. Les données devront être recueillies au fur et à mesure qu'elles sont obtenues, et transcrites dans ces cahiers de façon nette et lisible.
Les données erronées relevées sur les cahiers d'observation seront clairement barrées et les nouvelles données seront copiées, à côté de l'information barrée, accompagnées des initiales, de la date et éventuellement d'une justification par l'investigateur ou la personne autorisée qui aura fait la correction.
Les données sont recueillies sur un cahier d'observation papier.

1.32. SUIVI DE LA RECHERCHE

Le suivi de la recherche sera assuré par un technicien de recherche clinique. Il sera chargé, auprès de l'investigateur coordonnateur, de :

- La logistique et la surveillance de la recherche,
- L'établissement des rapports concernant son état d'avancement,

- La vérification de la mise à jour du cahier d'observation (demande d'informations complémentaires, corrections,…),
- La transmission des EIG au promoteur.

Il travaillera conformément aux procédures opératoires standardisées, en collaboration avec l'attaché de recherche clinique délégué par le promoteur.

1.33. CONTRÔLE DE QUALITÉ

Un attaché de recherche clinique mandaté par le promoteur visite de façon régulière chaque centre investigateur, lors de la mise en place de la recherche, une ou plusieurs fois en cours de recherche selon le rythme des inclusions et en fin de recherche. Lors de ces visites, les éléments suivants seront revus :

- consentement éclairé,
- respect du protocole de la recherche et des procédures qui y sont définies,
- qualité des données recueillies dans le cahier d'observation : exactitude, données manquantes, cohérence des données avec les documents source (dossiers médicaux, carnets de rendez-vous, originaux des résultats de laboratoire, etc,…),

Toute visite fera l'objet d'un rapport de monitorage par compte-rendu écrit transmis à l'investigateur du centre visité , à l'investigateur coordonnateur et au promoteur.

1.34. GESTION DES DONNÉES

Les données sont saisies en double saisie. La première saisie est réalisée sous EpiData, la validation est effectuée par un logiciel nommé DBS via un opérateur différent. Elle est réalisée par l'atelier de dactylo-codage de l'Université Bordeaux Segalen.

Les données sont validées conformément au plan de data management défini conjointement entre l'investigateur coordinateur et le Centre de Méthodologie et de Gestion des données. Les logiciels utilisés sont : ACCESS© et SAS®.

Le processus de gel/dégel des données est réalisé conformément à la procédure mise en place dans le Centre de Méthodologie et de Gestion des données (gel des données brutes au format XML et sous forme de table SAS).

1.35. AUDIT ET INSPECTION

Un audit peut être réalisé à tout moment par des personnes mandatées par le promoteur et indépendantes des responsables de la recherche. Il a pour objectif de s'assurer de la qualité de la recherche, de la validité de ses résultats et du respect de la loi et des règlementations en vigueur.

Les investigateurs acceptent de se conformer aux exigences du promoteur et à l'autorité compétente en ce qui concerne un audit ou une inspection de la recherche.

L'audit pourra s'appliquer à tous les stades de la recherche, du développement du protocole à la publication des résultats et au classement des données utilisées ou produites dans le cadre de la recherche.

CONSIDÉRATIONS ÉTHIQUES ET RÉGLEMENTAIRES

Le promoteur et l'investigateur s'engagent à ce que cette recherche soit réalisée en conformité avec la loi n°2004-806 du 9 août 2004, ainsi qu'en accord avec les Bonnes Pratiques Cliniques (I.C.H. version 4 du 1er mai 1996 et décision du 24 novembre 2006) et la déclaration d'Helsinki (qui peut être retrouvée dans sa version intégrale sur le site http://www.wma.net).

La recherche est conduite conformément au présent protocole. Hormis dans les situations d'urgence nécessitant la mise en place d'actes thérapeutiques précis, l' investigateur s'engage à respecter le protocole en tous points en particulier en ce qui concerne le recueil du consentement et la notification et le suivi des événements indésirables graves.

Cette recherche a reçu l'avis favorable du Comité de Protection des Personnes (CPP) SUD-OUEST et OUTRE MER III et l'autorisation de l'ANSM.

Le CHU de Bordeaux, promoteur de cette recherche, a souscrit un contrat d'assurance en responsabilité civile auprès de Gerling-Biomedicinsure conformément aux dispositions de l'article L1121-10 du code de la santé publique.

Les données enregistrées à l'occasion de cette recherche font l'objet d'un traitement informatisé à l'USMR, (responsable : Pr G Chene) dans le respect de la loi n°78-17 du 6 janvier 1978 relative à l'informatique, aux fichiers et aux libertés modifiées par la loi 2004-801 du 6 août 2004.

Cette recherche entre dans le cadre de la « Méthodologie de référence » (MR-001) en application des dispositions de l'article 54 alinéa 5 de la loi du 6 janvier 1978 modifiée relative à l'information, aux fichiers et aux libertés. Ce changement a été homologué par décision du 5 janvier 2006. Le CHU de Bordeaux a signé un engagement de conformité à cette « Méthodologie de référence ».

Cette recherche est enregistrée sur le site http://clinicaltrials.gov/.

AMENDEMENT AU PROTOCOLE

Toute modification substantielle, c'est à dire toute modification de nature à avoir un impact significatif sur la protection des personnes, sur les conditions de validité et sur les résultats de la recherche, sur la qualité et la sécurité des produits expérimentés, sur l'interprétation des documents scientifiques qui viennent appuyer le déroulement de la recherche ou sur les modalités de conduite de celle-ci, fait l'objet d'un amendement écrit qui est soumis au promoteur ; celui-ci doit obtenir, préalablement à sa mise en œuvre, un avis favorable du CPP et une autorisation de l'ANSM.

Les modifications non substantielles, c'est à dire celles n'ayant pas d'impact significatif sur quelque aspect de la recherche que ce soit, sont communiquées au CPP à titre d'information.

Tous les amendements sont validés par le promoteur, et par tous les intervenants de la recherche concernés, avant soumission au CPP et à l'ANSM. Cette validation peut nécessiter la réunion du CS.

Tous les amendements au protocole doivent être portés à la connaissance de tous les investigateurs qui participent à la recherche. Les investigateurs s'engagent à en respecter le contenu.

Tout amendement qui modifie la prise en charge des participants ou les bénéfices, risques et contraintes de la recherche fait l'objet d'une nouvelle note d'information et d'un nouveau formulaire de consentement dont le recueil suit la même procédure que celle précitée.

CONSERVATION DES DOCUMENTS ET DES DONNEES RELATIVES À LA RECHERCHE

Les documents suivants relatifs à cette recherche sont archivés conformément aux Bonnes Pratiques Cliniques et à la réglementation en vigueur :

- Par les médecins investigateurs :
  - **pour une durée de 15 ans suivant la fin de la recherche** :
    - Le protocole et les amendements éventuels au protocole
    - Les cahiers d'observation
    - Les dossiers source des participants ayant signé un consentement
    - Tous les autres documents et courriers relatifs à la recherche
  - **pour une durée de 30 ans suivant la fin de la recherche**
    - L'exemplaire original des consentements éclairés signés des participants

Tous ces documents sont sous la responsabilité de l'investigateur pendant la durée réglementaire d'archivage.

- Par le promoteur :
  - **pour une durée de 15 ans suivant la fin de la recherche :**
    - Le protocole et les amendements éventuels au protocole
    - L'original des cahiers d'observation

• Tous les autres documents et courriers relatifs à la recherche

- pour une durée de 30 ans suivant la fin de la recherche :

• Un exemplaire des consentements éclairés signés des participants

• Les documents relatifs aux événements indésirables graves

Tous ces documents sont sous la responsabilité du promoteur pendant la durée réglementaire d'archivage.

Aucun déplacement ou destruction ne pourra être effectué sans l'accord du promoteur. Au terme de la durée réglementaire d'archivage, le promoteur sera consulté pour destruction. Toutes les données, tous les documents et rapports pourront faire l'objet d'audit ou d'inspection.

RAPPORT FINAL

Dans un délai d'un an suivant la fin de la recherche (dernière visite du dernier patient) ou son interruption, un rapport final sera établi et signé par le promoteur et l'investigateur. Ce rapport sera tenu à la disposition de l'autorité compétente. Le promoteur transmettra à l'autorité compétente les résultats de la recherche sous forme d'un résumé du rapport final dans un délai d'un an après la fin de la recherche.

CONVENTION DE COOPÉRATION SCIENTIFIQUE ET REGLES RELATIVES À LA PUBLICATION

Cette recherche fera l'objet d'une convention de coopération scientifique entre les investigateurs, convention qui établira dans le détail des aspects de valorisation des résultats de la recherche.

1.36. COMMUNICATIONS SCIENTIFIQUES

L'analyse des données fournies par les centres investigateurs est réalisée par l'USMR,. Cette analyse donne lieu à un rapport écrit qui est soumis au promoteur, qui transmettra au Comité de Protection des Personnes et à l'autorité compétente.

Toute communication écrite ou orale des résultats de la recherche doit recevoir l'accord préalable de l'investigateur coordonnateur et, le cas échéant, de tout comité constitué pour la recherche.

La publication des résultats principaux mentionne le nom du promoteur, de tous les investigateurs ayant inclus ou suivi des participants dans la recherche, des méthodologistes, biostatisticiens et data managers ayant participé à la recherche, des membres du (des) comité(s) constitué(s) pour la recherche et la source de financement. Il sera tenu compte des règles internationales d'écriture et de publication (Convention de Vancouver, février 2006).

1.37. COMMUNICATION DES RÉSULTATS AUX PARTICIPANTS

Conformément à la loi n°2002-303 du 4 mars 2002, les participants sont informés, à leur demande, des résultats globaux de la recherche.

1.38. CESSION DES DONNÉES

Le recueil et la gestion des données sont assurés par l'USMR. Les conditions de cession de tout ou partie de la base de données de la recherche sont décidées par le promoteur de la recherche et font l'objet d'un contrat écrit.

RÉFÉRENCES BIBLIOGRAPHIQUES

Abel LA, Unverzagt F, Yee RD. Effects of stimulus predictability and interstimulus gap on saccades in Alzheimer's disease. Dement Geriatr Cogn Disord. 2002; 13: 235-243.

Amieva H, Le Goff M, Millet X, Orgogozo JM, Pérès K, Barberger-Gateau P, Jacqmin-Gadda H, Dartigues JF. Prodromal Alzheimer's disease: successive emergence of the clinical symptoms. Ann Neurol. 2008; 64: 492-498.

Andrieu S, Coley N, Aisen P, Carrillo MC, Dekosky S, Durga J, Fillit H, Frisoni GB, Froelich L, Gauthier S, Jones R, Jönsson L, Khachaturian Z, Morris JC, Orgogozo JM, Ousset PJ, Robert P, Salmon E, Sampaio C, Verhey F, Wilcock G, Vellas B. Methodological issues in primary prevention trials for neurodegenerative dementia. J Alzheimer's Dis 2009; 16 : 235-270.

Barker RA, Michell AW. "The eyes have it". Saccadometry in Parkinson's disease; Exp Neurol 2009; 07015.

Barbeau E, Didic M, Tramoni E, Felician O, Joubert S, Sontheimer A, Ceccaldi M, Poncet M. Evaluation of visual recognition memory in MCI patients. Neurology. 2004; 62: 1317-1322.

Boxer AL, Garbutt S, Rankin KP, Hellmuth J, Neuhaus J, Miller BL, Lisberger SG. Medial versus lateral frontal lobe contributions to voluntary saccade control as revealed by the study of patients with frontal lobe degeneration. J Neurosci. 2006; 26: 6354-6363.

Brown MR, Vilis T, Everling S. Frontoparietal activation with preparation for antisaccades. J Neurophysiol. 2007; 1751-1762.

Crawford TJ, Higham S, Renvoize T, Patel J, Dale M, Suriya A, Tetley S. Inhibitory control of saccadic eye movements and cognitive impairment in Alzheimer's disease. Biol Psychiatry. 2005; 57: 1052-1060.

Daffner KB, Scinto LFM, Wentraub S, Guinessey JE, Mesulam MM. Diminisher curiosity in patients with probable Alzheimer's disease as measured by exploratory eye movements. Neurology 1992; 42: 320-328.

Deloche G, Hannequin D, Dordain M, Metz-Lutz MN, Kremin H, Tessier C, Vendrell J, Cardebat D, Perrier D, Quint S, Pichard B. Diversity of patterns of improvement in confrontation naming rehabilitation: some tentative hypotheses. J Commun Disord. 1997; 30: 11-21.

Dubois B, Feldman HH, Jacova C, Dekosky ST, Barberger-Gateau P, Cummings J, Delacourte A, Galasko D, Gauthier S, Jicha G, Meguro K, O'brien J, Pasquier F, Robert P, Rossor M, Salloway S, Stern Y, Visser PJ, Scheltens P. Research criteria for the diagnosis of Alzheimer's disease: revising the NINCDS-ADRDA criteria. Lancet Neurol. 2007; 6: 734-746.

Dubois B, Albert ML. Amnestic MCI or prodromal Alzheimer's disease? Lancet Neurol. 2004; 3: 246-248.

Dubois B, Feldman HH, Jacova C, Cummings JL, Dekosky ST, Barberger-Gateau P, Delacourte A, Frisoni G, Fox NC, Galasko D, Gauthier S, Hampel H, Jicha GA, Meguro K, O'Brien J, Pasquier F, Robert P, Rossor M, Salloway S, Sarazin M, de Souza LC, Stern Y, Visser PJ, Scheltens P.Revising the definition of Alzheimer's disease: a new lexicon. Lancet Neurol. 2010 ; 9: 1118-11127.

Fabrigoule C, Rouch I, Taberly A, Letenneur L, Commenges D, Mazaux JM, Orgogozo JM, Dartigues JF. Cognitive process in the preclinical phase of dementia. Brain 1998; 12: 135-141.

Garbutt S, Matlin A, Hellmuth J, Schenk AK, Johnson JK, Rosen H, Dean D, Kramer J, Neuhaus J, Miller BL, Lisberger SG, Boxer AL. Oculomotor function in frontotemporal lobar degeneration, related disorders and Alzheimer's disease. Brain. 2008; 131: 1268-1281.

Grosbras MH, Lobel E, Berthoz A. [The control of gaze (2): cortical control of ocular saccades: functional brain imaging data] Med Sci (Paris). 2004; 20: 225-230.

Hughes CP, Berg L, Danziger WL, Coben LA, Martin RL. A new clinical scale for the staging of dementia. Br J Psychiatry. 1982; 140: 566-572.

Korf ES, Wahlund LO, Visser PJ, Scheltens P.Medial temporal lobe atrophy on MRI predicts dementia in patients with mild cognitive impairment. Neurology. 2004; 63: 94-100.

Lawton MP, Brody EM. Assessment of older people: self-maintaining and instrumental activities of daily living. Gerontologist. 1969; 9: 179-186.

Loughland CN, Williams LM, Gordon E. Schizophrenia and affective disorder show different visual scanning behavior for faces: a trait versus state-based distinction ? Biol Psychiatry 2002; 52: 338-348.

Mosimann UP, Felblinger J, Ballinari P, Hess CW, Müri RM. Visual exploration behaviour during clock reading in Alzheimer's disease. Brain. 2004; 127: 431-438.

Mosimann UP, Müri RM, Burn DJ, Felblinger J, O'Brien JT, McKeith IG. Saccadic eye movement changes in Parkinson's disease dementia and dementia with Lewy bodies. Brain. 2005; 128: 1267-1276.

Okello A, Koivunen J, Edison P, Archer HA, Turkheimer FE, Någren K, Bullock R, Walker Z, Kennedy A, Fox NC, Rossor MN, Rinne JO, Brooks DJ. Conversion of amyloid positive and negative MCI to AD over 3 years: an 11C-PIB PET study. Neurology. 2009; 73: 754-760.

Petrie EC, Cross DJ, Galasko D, Schellenberg GD, Raskind MA, Peskind ER, Minoshima S. Preclinical evidence of Alzheimer changes: convergent cerebrospinal fluid biomarker and fluorodeoxyglucose positron emission tomography findings. Arch Neurol 2009; 66: 632-637.

Puel M, Hugonot-Diener L.[Presentation by the GRECO group of the French adaptation of a cognitive assessment scale used in Alzheimer type dementia]. Presse Med. 1996; 25: 1028-1032.

Reitan RM. Investigation of relationships between psychometric and biological intelligence. J Nerv Ment Dis. 1956; 123: 536-541.

Rivaud-Péchoux S, Vidailhet M, Brandel JP, Gaymard B. Mixing pro- and antisaccades in patients with parkinsonian syndromes.Brain. 2007; 130: 256-264.

Roman GC, Sachdev P, Royall DR, Bullock RA, Orgogozo JM et al. Vascular cognitive disorder: a new diagnostic category updating vascular cognitive impairment and vascular dementia. J Neurol Sci 2004; 226: 81-87.

Rösler A, Mapstone M, Hays-Wicklund A, Gitelman DR, Weintraub S. The "zoom lens" of focal attention in visual search: changes in aging and Alzheimer's disease. Cortex. 2005; 41: 512-519.

Rösler A, Mapstone ME, Hays AK, Mesulam MM, Rademaker A, Gitelman DR, Weintraub S. Alterations of visual search strategy in Alzheimer's disease and aging. Neuropsychology. 2000; 14: 398-408.

Rüb U, Del Tredici K, Schultz C, Büttner-Ennever JA, Braak H. The premotor region essential for rapid vertical eye movements shows early involvement in Alzheimer's disease-related cytoskeletal pathology. Vision Res. 2001; 41: 2149-2156.

Scheinin NM, Aalto S, Koikkalainen J, Lötjönen J, Karrasch M, Kemppainen N, Viitanen M, Någren K, Helin S, Scheinin M, Rinne JO. Follow-up of [11C]PIB uptake and brain volume in patients with Alzheimer disease and controls. Neurology. 2009; 73: 1186-1192.

Scheltens P, Leys D, Barkhof F, Huglo D, Weinstein HC, Vermersch P, Kuiper M, Steinling M, Wolters EC, Valk J.Atrophy of medial temporal lobes on MRI in "probable" Alzheimer's disease and normal ageing: diagnostic value and neuropsychological correlates. J Neurol Neurosurg Psychiatry. 1992; 55: 967-972.

Shafiq-Antonacci R, Maruff P, Masters C, Currie J. Spectrum of saccade system function in Alzheimer disease.Arch Neurol. 2003; 60: 1272-1278.

Singh NN, Ellis CR, Wechsler H. Psychopharmacoepidemiology of mental retardation: 1966 to 1995. J Child Adolesc Psychopharmacol. 1997; 7: 255-266.

Van der Linden M, Juillerat AC. Memory systems and memory disorders. Rev Prat. 2003; 53: 400-405.

Yesavage JA, Brink TL, Rose TL, Lum O, Huang V, Adey M, Leirer VO. Development and validation of a geriatric depression screening scale: a preliminary report. J Psychiatr Res. 1982-1983; 17: 37-49.

Résumé

Les maladies neurodégénératives avec démence constituent un réel problème de santé publique qui augmente avec le vieillissement de la population dans les pays développés. En effet, environ 860 000 personnes souffrent de démence de type Alzheimer chez des sujets âgés de plus de 65 ans dans la population française. Le ministère tunisien de la santé publique a signalé environ 20 000 cas confirmés. Un diagnostic précoce et non invasif de la démence est primordial. Diverses expériences ont été réalisées afin d'identifier les premiers symptômes de la démence. L'évaluation oculomotrice est l'une d'entre elles.

Avant de faire des expériences avec l'inclusion des patients atteints des maladies neurodégénératives, il est important de mesurer l'impact de l'attention visuelle des sujets témoins normaux sur des dégradations générées dans des vidéos naturelles. Dans ce cadre se situent les travaux de la thèse. Initialement, une étude bibliographique approfondie a été élaborée à propos des modèles de prédiction de l'attention visuelle et l'apprentissage automatique profond pour prédire de la saillance . Les différents modèles de prédiction de la saillance visuelle proposés dans la littérature et les principales techniques utilisées ont vigoureusement étudié. À partir d'une réflexion approfondie et une analyse critique des travaux existants, nous avons fixé les objectifs de la thèse ainsi que les contributions autour desquelles s’inscrit le travail de recherche. Les contributions consistent principalement à concevoir un modèle complet de prédiction de saillance visuelle. Dans cette thèse, nous proposons de modéliser l'efficacité des vidéos par des outils d'apprentissage automatique en particulier avec les réseaux de convolution profonde. Un premier modèle permet d'insérer l'information de mouvement résiduel côte à côte avec la valeur RVB pour chaque image de la vidéo. Ce modèle, appelé "ChaboNet4k", classifie les informations d'entrée (mouvement résiduel et valeurs RVB) en deux classes: (saillantes et non-saillantes).

Trois objectifs principaux sont le but de cette thèse. Tout d'abord, nous souhaitons mieux comprendre les processus attentionnels qui guident le regard vers

des régions particulières du champ visuel. Deuxièmement, modéliser ces processus par des outils d'apprentissage automatique en vue de leurs succès approuvés. Et enfin d'appliquer ce modèle de prédiction de saillance pour tester les patients atteints des maladies neurodégénératives. La modélisation de l'attention visuelle avec les réseaux de convolution profond nous permettra de combiner les caractéristiques de bas niveau avec ceux de haut niveau pour la prédiction des régions saillants par un ensemble de sujets lors de l'affichage d'une vidéo naturelle. Plusieurs travaux et modèles conçus pour l'attention visuelle avec l'apprentissage profond existent concernant l'image statique mais encore assez peu qui se proposent d'étudier les vidéos. Nous serons intéressés par ce travail dans l'étude de scènes dynamiques à travers une base de données de vidéos. Nous adapterons deux approches complémentaires. La première permet de désigner l'architecture du réseau de convolution profond assurant la prédiction des zones saillantes. Nous définissons le problème de l'apprentissage comme le problème de classifi- cation de deux classes (saillant, non saillant) en se référant aux enregistrements des mouvements oculaires des sujets visualisant des vidéos naturelles avec des contenus variés. La deuxième approche de l'apprentissage par transfert nous permettra de proposer un modèle inspiré de l'architecture déjà désignée pour résoudre le problème des ensembles de données disponibles avec peu de vidéos.

Dans ce travail, nous nous sommes intéressés à comprendre les processus attentionnels impliqués dans la perception d'une scène visuelle et à comprendre la fonction des réseaux de convolution profonde et enfin à proposer une architecture originale pour la prédiction de la saillance. Dans cette étude, nous avons considéré le cas des vidéos. Pour explorer une vidéo, nous concentrons notre attention sur certaines régions saillantes dont le mouvement et l'aspect sémantique de l'objet d'intérêt représentent un rôle majeur. Dans cette thèse, nous avons proposé une architecture de réseaux de convolution profonds acceptant comme entrée la fonction de mouvement résiduel et permettant de prédire les régions saillants d'une vidéo. Et, nous avons étudié l'apprentissage par transfert pour un réseau de convolution profond sur des vidéos spécifiques réalisées pour étudier les maladies neurodégénératives. Les points de fixations humaines sur ce type de vidéos ont été calculés à travers des expériences psychovisuelles. Nous proposons de faire le point sur les différents résultats obtenus et de présenter plusieurs perspectives que nous considérons pertinentes pour poursuivre ce travail.

Nous avons proposé notre solution de prédiction des cartes de saillance en vidéo dans le cadre de l'apprentissage en profondeur. Il consiste en deux étapes. Tout d'abord, un réseau de convolution profond a été conçu pour prédire les zones saillants (patches) dans le contenu vidéo. Ensuite, les cartes denses de saillance visuelle prédites ont été calculées sur la base des résultats de classification de patch. Nous avons construit une architecture CNN profonde adéquate sur la base de Caffe CNN. Comme les CNN profonds sont sensibles au bruit dans les données d'entraînement, nous avons proposé une solution adaptée pour le réduire. Dans notre cas de prédiction des patchs saillants, les règles de production vidéo telles que la règle des tiers qui est utilisée pour la sélection des patchs non saillants, ont permis d'augmenter la précision. Bien que la recherche sur l'état de l'art n'utilise que les valeurs primaires RVB pour la prédiction dans le contenu visuel, nous avons montré que pour la vidéo, l'ajout de caractéristiques exprimant la sensibilité du système visuel humain au mouvement résiduel est important. Nous avons comparé les performances de prédiction avec des CNN profonds lorsque différents types de caractéristiques ont été ingérés (telles que les valeurs de pixels de couleur seulement, ou les valeurs de couleur avec un mouvement résiduel).

Nous proposons une architecture Deep CNN relativement peu profonde et nous l'avons comparée à architectures similaires AlexNet et LeNet. Notre architecture a montré une meilleure puissance de prédiction en termes de précision moyenne et de stabilité de la phase d'entraînement.

Nous avons répondu au problème de la thèse. Tout d'abord, un modèle à quatre canaux basé sur les couleurs R, V, B et le mouvement a été proposé. Ensuite, ce modèle a été enrichi avec sept autres canaux résumant les différents types de contraste déjà étudiés pour la prédiction de saillance. Tout au long de cette thèse, nous avons utilisé des bases de données qui recueillent des informations sur le regard humain enregistré à travers des expériences psychovisuelles. L'utilisation de cette information nous a permis de définir la classe cible des régions saillants. Les principales contributions de la thèse sont:

- ChaboNet: une couche de données d'entrée spécifique de CNN profond pour la prédiction de saillance visuelle. Ici, nous présentons la première contribution principale concernant la modélisation de la saillance visuelle pour les vidéos naturelles. Tout d'abord, l'utilisation de la carte dense de fixation des yeux dans

l'entraînement des modèles profonds CNNs assure la combinaison des deux domaines bottum-up et top-down. Deuxièmement, pour le traitement vidéo, les signaux temporels sont principalement utilsés pour détecter la région saillant. Les expériences ont montré des résultats prometteurs. Nous avons mené quelques expériences en utilisant sept types de contrastes comme entrée du CNN profond afin d'exploiter la sensibilité de la prédiction de saillance sur les contrastes. Cela nous a permis d'être sûr de certains choix du modèle et de limiter le modèle d'entrée sur la composante temporelle avec les valeurs RVB.

• Proposition d'une méthode inspirée de la méthode du Wooding pour la génération des cartes dense de saillance à partir des réponses de probabilité du modèle du réseau profond entraîné. Génération de la carte de saillance ayant la même taille de trames d'entrée: nous proposons une méthode originale et spécifique pour générer la carte de saillance finale. Les codes ont été optimisés avec un algorithme parallèle qui réduit le temps de génération des cartes de saillance.

• L'étude de l'effet du bruit de données sur l'apprentissage des réseaux profonds (entraînement et prédiction) : une étude et des expériences sur les réseaux de convolution profonds ont été menées afin de mettre en évidence la sensibilité de ce type de réseaux au bruit dans les données d'entraînement. Les résultats montrent qu'avec des données pures, le modèle prédit avec une précision plus intéréssante.

• Transférer l'apprentissage avec CNN profond pour la prédiction de saillance : Proposition d'une méthode d'apprentissage par transfert pour résoudre le problème de taille des jeux de données disponibles. La deuxième contribution principale de la thèse est l'utilisation du modèle déjà entraîné sur un grand ensemble de données pour affiner un jeu de données plus petit. Cette méthode a été étudiée et expérimentée sur quatre petits ensembles de données.

• Création de vidéos d'une base de données spécifiques pour le teste des patients atteints de maladies neuro-dégénératives et l'application d'une méthode d'apprentissage par transfert pour la base de données spécifique créée.

• Des expériences de suivi des yeux: une expérience de suivi des yeux est

proposée pour la base de données GTEA et nos vidéos créées pour tester les patients atteints de maladies neurodégénératives. Cette expérience a été utilisée pour enregistrer les positions des yeux de sujets sains pendant qu'ils regardaient librement des vidéos. Ces données oculométriques nous ont permis, d'abord, d'étudier et d'évaluer notre architecture proposée avec des réseaux CNN profond pour la prédiction de saillance. Deuxièmement, il assure la définition des zones saillants pour lancer l'entraînement de modèle de l'architecture proposée. Nous avons pu tirer plusieurs conclusions telles que les sujets normaux sont attirés par des objets significatifs. En effet, ils poursuivent les nouveaux objets en laissant tomber la dégradation.

Les approches proposées dans le cadre des contributions ont été évalués et comparé à d'autre travaux de l'état de l'art.

Ce travail ouvre de nombreuses perspectives qui peuvent être envisagées soit comme une amélioration ou une extension directe de ce travail, soit comme une nécéssité d'études plus approfondies à long terme. Une exploitation plus poussée des possibilités du modèle peut être réalisée en amplifiant le ChaboN et4k avec une étape de « fine-tuning » d'un autre modèle entraîné en particulier par une opération « net surgery ». L'utilisation d'autres architectures de réseaux de convolution profonds qui modélise un modèle en temps réel de prédiction de saillance sur des vidéos naturelles ou sur des vidéos égocentriques, présente une nouvelle piste de recherche. En conclusion, nous pensons que le modèle de saillance proposé utilisant des CNN profonds a une très bonne perspective d'application, en particulier dans les diagnostics neurodégénératifs et plusieurs autres applications de prédiction de saillance telles que la compression de la vidéo. Un total de dix logiciels et scripts ont été développés pour ce projet de recherche.