

Modélisation spatio-temporelle de l'attention visuelle

O. Le Meur^{1,2}

P. Le Callet¹

D. Barba¹

D. Thoreau²

¹ THOMSON

1, Avenue de Belle-Fontaine - BP 19

35511 Cesson-Sévigné

France

² IRCCyN

Rue Christian Pauc - BP 50609

44306 Nantes

France

Résumé

Le système visuel, intrinsèquement limité, utilise des mécanismes bien particuliers pour produire une représentation dite économique de notre environnement visuel. De type endogène ou exogène, l'attention visuelle fait partie des mécanismes actifs de sélection et de hiérarchisation de l'information. Dans cette contribution, nous proposons de détailler une modélisation de l'attention visuelle exogène. L'objectif est donc de déterminer les zones d'une image ou d'une séquence d'images qui attirent le regard. Ces zones sont communément qualifiées d'intérêt. L'évaluation de cette modélisation est effectuée à partir d'une référence acquise via des expérimentations oculométriques. Deux comparaisons sont réalisées. La première concerne l'évaluation des performances du modèle dans la dimension spatiale. Les résultats obtenus sont bien meilleurs que ceux obtenus par un autre modèle, considéré aujourd'hui comme une référence. La seconde concerne l'évaluation du modèle dans la dimension spatio-temporelle. A partir d'une classification supervisée, on montre qu'en moyenne 77% des pixels sont correctement classés.

Mots clefs

attention visuelle exogène, système visuel, expérimentations oculométriques.

1 Introduction

Bien que l'environnement visuel dans lequel nous évoluons soit constitué d'une quantité d'information indénombrable, notre système visuel est capable d'appréhender et d'interpréter avec précision l'ensemble de ces informations visuelles. Des mécanismes particuliers ainsi que des stratégies d'exploration de l'espace visuel sont nécessaires pour résoudre cette situation paradoxale. Les premiers permettent de construire une représentation économique du contenu visuel. Cette représentation, où la redondance a été supprimée, est précise au centre de la rétine (la fovéa) et grossière dans la périphérie. Concernant les stratégies d'explorations, elles sont au nombre de deux : l'attention endogène et l'attention exogène. La première, également appelée *Top-Down*, est une stratégie pilotée

par la tâche que nous avons à effectuer, impliquant un contrôle volontaire et cognitif des mouvements oculaires. Ce mécanisme, nécessitant toutes les ressources attentionnelles, est déployé pour effectuer une tâche : reconnaître un lieu sur une photo, chercher l'homme portant une casquette verte... La stratégie exogène, plus communément appelée *Bottom-Up*, permet, quant à elle, de sélectionner les informations visuelles selon leur saillance. Ce type de stratégie fait référence à l'attention visuelle involontaire, c'est à dire à un traitement automatique très rapide réalisé inconsciemment. Ce sont les caractéristiques de notre champ visuel qui attirent notre regard.

La modélisation de l'attention visuelle est un véritable enjeu, aussi bien économique qu'intellectuel. Économique, car les applications sont nombreuses. Les plus directes et les plus intéressantes pour notre étude sont la compression vidéo, le tatouage numérique ou encore l'évaluation de qualité. Intellectuel, car une modélisation pertinente de l'attention visuelle est le point de convergence de nombreux domaines d'études : la neurobiologie, la physiologie, la psychophysiques et bien entendu le domaine du traitement d'images en sont des exemples.

La modélisation spatio-temporelle de l'attention exogène, c'est à dire la détection des zones de saillance d'une séquence d'images, doit permettre de construire une carte de saillance associée à chaque image. Le concept de la carte de saillance, encodant spatialement le degré d'intérêt de chaque pixel, a été introduit par Koch et Ullman [?]; ces derniers sont également à l'origine d'une architecture, biologiquement plausible, aujourd'hui devenue majeure pour la modélisation de l'attention visuelle.

Basée sur cette architecture, cette contribution présente un modèle spatio-temporel de l'attention visuelle exogène, conçu à partir de nombreuses propriétés du système visuel. Dans la seconde partie, les grandes caractéristiques du modèle sont décrites. Afin d'évaluer la pertinence du modèle, une comparaison est réalisée avec la vérité terrain. La troisième partie présente la détermination de cette vérité terrain issue d'expérimentations oculométriques. La comparaison des prédictions et des observations est effectuée dans la quatrième partie.

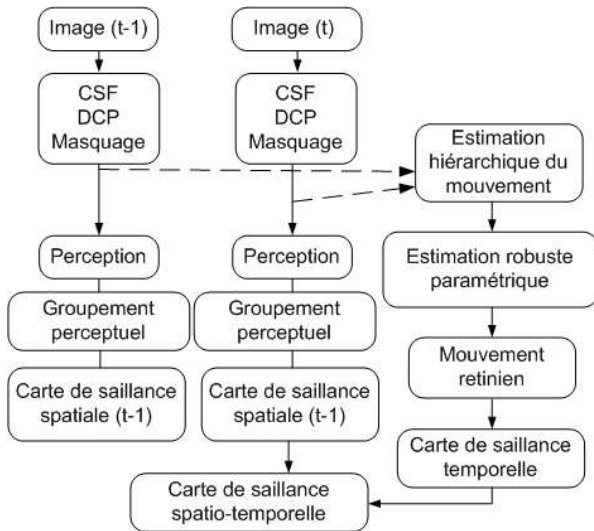


Figure 1 – Synoptique de la modélisation proposée

2 Modèle spatio-temporel de l'attention visuelle exogène

Le synoptique de l'approche proposée est donné à la figure ???. Les grandes lignes du modèles sont détaillées brièvement dans les paragraphes suivants. La partie spatiale est d'abord décrite suivie de la partie temporelle. Le lecteur pourra se référer aux articles [?, ?, ?] pour de plus amples détails.

2.1 Modélisation spatiale

La modélisation spatiale de l'attention visuelle est composée de trois parties séquentielles.

La première partie regroupe les outils modélisant le fait que notre système visuel n'apprécie pas de la même façon les composantes visuelles de notre environnement. Cette sensibilité limitée est simulée par l'utilisation de fonctions de sensibilité aux contrastes (abrégé CSF pour l'acronyme anglais *Contrast Sensitivity Function*) [?] et de l'utilisation du masquage visuel intra et inter composantes [?]. Ces fonctions sont appliquées aux composantes (A, Cr_1, Cr_2) de l'espace de couleurs antagonistes de Krauskopf, déduites des composantes RGB d'une image. Une décomposition hiérarchique en canaux perceptuels, notée DCP sur la figure ?? et donnée à la figure ??, simule le pavage fréquentiel du système visuel [?]. A partir du spectre de fréquence, un ensemble de sous bandes ayant une gamme de fréquences radiales et une sélectivité angulaire particulière est définie. Chaque sous bande peut être en fait considérée comme l'image neuronale délivrée par une population de cellules visuelles réagissant à un ensemble de fréquences et d'orientations [?]. L'utilisation de CSF et de masquage permet d'obtenir un espace psychovisuel dans lequel toutes les données (achromatiques et chromatiques) s'expriment en fonction de leur degré de visibilité. Par conséquent, le problème majeur des

modèles de l'état de l'art qui dénaturent les dynamiques des signaux en utilisant un opérateur normalisation global est résolu.

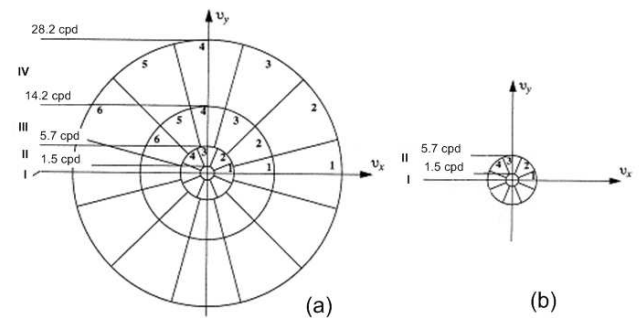


Figure 2 – Décomposition en canaux perceptuels : (a) décomposition de la composante achromatique en 17 sous bandes réparties sur les couronnes I à IV ; (b) décomposition des composantes chromatiques en 5 sous bandes réparties sur les couronnes I à II.

La seconde partie concerne le mécanisme de la perception visuelle. Les valeurs de visibilité des données de l'espace psychovisuel sont donc transformées en saillance afin d'extraire les caractéristiques visuelles portant de l'information importante. Les transformations utilisées conduisent alors à la création d'une représentation économique de notre environnement. L'organisation des champs récepteurs des cellules visuelles, que ce soit rétiniennes ou corticales répond tout à fait à ce besoin. Ces derniers sont circulaires, ayant une direction préférée (pour les cellules corticales) et sont constitués d'un centre et d'un pourtour ayant des réponses antagonistes. Cette organisation leur confère donc la propriété de répondre fortement sur les contrastes et de ne pas répondre sur les zones uniformes. La modélisation de ce type de cellules s'effectue via des différences de Gaussiennes (DoG) orientées ou non. Le profil du filtre utilisé est donné à la figure ??.

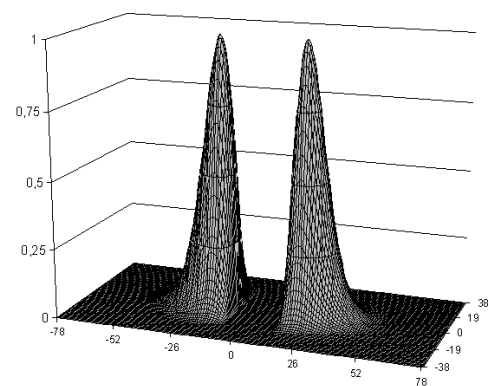


Figure 3 – Profil de la fonction modélisant la contribution inhibitrice d'une cellule corticale.

Ainsi, les sous bandes provenant des trois composantes sont convoluées avec un opérateur proche d'une DoG. La perception consiste également à accentuer certaines caractéristiques essentielles à l'interprétation de l'information. En suivant les principes de l'école Gestaltiste, un filtre en papillon [?] est appliqué afin de renforcer les contours co-linéaires, alignés et de faible courbure. Les principes Gestaltiens de bonne continuité et de co-linéarité sont donc utilisés.

Finalement, afin de construire la carte de saillance spatiale, une fusion des différentes composantes est effectuée groupant ou liant des éléments, a priori indépendants, pour former une structure compréhensible par le cerveau. La fusion est basée sur deux mécanismes :

- une compétition intra permettant d'identifier les zones les plus pertinentes de la densité ;
- une compétition inter cartes tirant profit de la redondance et de la complémentarité des différentes densité. L'utilisation de la redondance inter cartes permet de renforcer la saillance de certaines zones lorsque celles-ci génèrent de la saillance dans plusieurs dimensions. Par contre, lorsqu'une zone ne génère de la saillance que dans une seule dimension visuelle, il est nécessaire d'utiliser la complémentarité inter cartes.

La fusion cohérente est présentée pour deux cartes, notées DS^{C_1} et DS^{C_2} issues d'une composante C_1 et C_2 . La généralisation à n densités est facilement envisageable. La densité finale, notée DS , est obtenue par la fusion des cartes DS^{C_1} et DS^{C_2} , notée $\mathcal{F}(DS^{C_1}, DS^{C_2})$:

$$DS(s) = \mathcal{F}(DS^{C_1}(s), DS^{C_2}(s)) \quad (1)$$

L'opérateur de fusion $\mathcal{F}(\cdot)$ est composé d'une série de trois transformations que nous allons décrire. Ces trois transformations s'utilisent successivement.

Une étape de normalisation

Tout d'abord, un procédé de fusion ne peut se faire sans une étape préalable de normalisation de dynamique. Contrairement aux procédés de fusion proposés par L. Itti [?] qui utilisaient une normalisation à partir du maximum global de chaque carte, la normalisation que nous utilisons se base sur le maximum empirique de chaque dimension visuelle. Ces maximums sont déterminés expérimentalement en utilisant des tests particuliers. Par exemple, pour la composante Cr_1 , une image à luminance uniforme mais présentant un motif rouge saturé va générer une dynamique proche de la dynamique maximale de l'axe visuel Cr_1 . La répétition de ce type d'expérimentation a permis de définir les maximums empiriques des composantes A , Cr_1 et Cr_2 .

Les deux cartes de densité DS^{C_1} et DS^{C_2} sont donc normalisées pour être sur la même dynamique. Ensuite, afin de construire un histogramme, ces données sont quantifiées linéairement sur L niveaux. Elles sont respectivement notées $DS_{NQ}^{C_1}$ et $DS_{NQ}^{C_2}$.

Compétition intra carte

La compétition intra carte modifie la valeur de chaque site s des cartes $DS_{NQ}^{C_1}$ et $DS_{NQ}^{C_2}$ en fonction de la valeur du maximum local le plus proche. Ce type de compétition est donné par la relation suivante :

$$intraMap^{C_1}(s) = \frac{DS_{NQ}^{C_1}(s)}{PlusProcheMax_{C_1}(s)} \quad (2)$$

$$intraMap^{C_2}(s) = \frac{DS_{NQ}^{C_2}(s)}{PlusProcheMax_{C_2}(s)} \quad (3)$$

La fonction $PlusProcheMax_{C_1}$ (respectivement $PlusProcheMax_{C_2}$) retourne la valeur du maximum local de la composante C_1 (respectivement C_2) la plus proche de la valeur du site s . Cette valeur est extraite de la liste \mathcal{L}_1 (respectivement \mathcal{L}_2) de taille k_1 (respectivement k_2) valeurs. La taille des listes est déterminée de façon à avoir un rapport entre le maximum local n et le maximum local $n + 1$ supérieur à un seuil, fixé arbitrairement à 1.3. Cet artifice permet de prendre uniquement en compte les principales zones de saillance. Par ailleurs, le maximum local $n + 1$ est déterminé en inhibant une zone circulaire centrée autour du maximum local n et d'un rayon de un degré visuel, reproduisant une sélection de type *Winner-Take-All*.

Compétition inter cartes

La compétition inter cartes tire profit de la redondance et de la complémentarité des différentes cartes. Le terme *interMap*, lié à la compétition inter cartes, est donné par la relation suivante :

$$interMap(s) = complementarite(s) + redondance(s) \quad (4)$$

La complémentarité, notée *complementarite* dans la relation (??) s'obtient en sommant les résultats de la compétition intra carte :

$$complementarite(s) = intraMap^{C_1}(s) + intraMap^{C_2}(s) \quad (5)$$

La redondance inter cartes est traitée à partir d'une analyse conjointe des distributions des cartes à fusionner. Elle est notée *redondance* et donnée par la relation (??) :

$$redondance(s) = intraMap^{C_1}(s) \times intraMap^{C_2}(s) \times \alpha \quad (6)$$

avec, $\alpha = \frac{\text{Log} \frac{N}{H(DS_{NQ}^{C_1}(s), DS_{NQ}^{C_2}(s))}}{3 \text{Log}(L)}$ et N le nombre de sites des cartes considérées.

Le facteur α déduit de l'histogramme conjoint des cartes $DS_{NQ}^{C_1}$ et $DS_{NQ}^{C_2}$ modifie la valeur du site s considéré en fonction de sa probabilité d'apparition. Cette approche statistique est inspirée des travaux de A. Oliva [?] et de ceux de B. Bruce [?]. Ces travaux utilisent le fait que la quantité d'informations portée par un site s est inversement proportionnelle à sa probabilité d'apparition. Par conséquent, le facteur déduit de l'analyse conjointe augmente la valeur

d'un site s lorsque sa probabilité d'apparition est faible. Réciproquement, la valeur du site s est diminuée lorsque sa probabilité d'apparition est forte.

L'opérateur de fusion \mathcal{F} est donc équivalent au terme *interMap*. Ce dernier intègre à la fois la compétition intra carte et la compétition inter cartes.

2.2 Modélisation temporelle

Dans un contexte animé, les contrastes de mouvement sont certainement les attracteurs visuels les plus significatifs. Il est clair qu'un objet en déplacement sur un fond fixe, ou réciproquement un objet fixe sur un fond mouvement, attire l'attention visuelle. Pour déterminer ces contrastes, la prise en compte des mouvements oculaires de poursuite est primordiale. Ces mouvements oculaires permettent de compenser naturellement le déplacement d'un objet. La vélocité du mouvement considéré, exprimée dans le référentiel rétinien est alors quasi nulle. Pour déterminer les contrastes de mouvement les plus pertinents, il est par conséquent nécessaire de compenser le mouvement inhérent de la caméra, supposé dominant. A partir d'un champ de vecteurs, issu d'un estimateur de mouvement hiérarchique local travaillant sur la décomposition en canaux perceptuels, un modèle paramétrique affine complet est calculé grâce à la technique d'estimation robuste proposée par Odohez [?]. Le mouvement rétinien correspond alors à la différence entre le mouvement local \vec{V}_{local} et le mouvement dominant \vec{V}_{Θ} (θ contient les 6 paramètres du modèle de mouvement affine) :

$$\vec{V}_{relatif}(s) = \vec{V}_{\Theta}(s) - \vec{V}_{local}(s) \quad (7)$$

La relation (??) est modifiée afin de prendre en compte la vélocité maximale théorique du mouvement oculaire de poursuite. Plus la vélocité du mouvement relatif est supérieure à la vélocité maximale de poursuite et plus la saillance doit être atténuée. Par conséquent, lorsque $\|\vec{V}_{relatif}(s)\| > \vec{v}_{max}$, on a :

$$\vec{V}_{relMod}(s) = \vec{V}_{relatif}(s) \cdot \left\{ \frac{\vec{v}_{max}}{\|\vec{V}_{relatif}(s)\|} \right\}^{\gamma} \quad (8)$$

avec, \vec{v}_{max} la vélocité maximale de poursuite de l'oeil. Le paramètre γ contrôle la modification de la pente. En pratique, nous avons γ égal à 3.

Une autre propriété est également à considérer. Il est relativement connu qu'un objet en mouvement sur un fond fixe attire plus facilement l'attention qu'un objet fixe sur un fond en mouvement. L'amplitude du mouvement dominant est donc un facteur important pour déterminer la saillance finale. Une façon pertinente pour l'évaluer consiste à prendre la valeur médiane de l'histogramme des mouvements relatifs quantifiées, noté par $Med(\|\vec{V}_{relMod}(s)\|_Q)$.

La saillance temporelle S^T est alors déduite en pondérant le mouvement relatif quantifié par

$Med(\|\vec{V}_{relMod}(s)\|_Q)$:

$$S^T(s) = \frac{\|\vec{V}_{relMod}(s)\|_Q}{Med(\|\vec{V}_{relMod}(s)\|_Q)} \quad (9)$$

La carte de saillance S^T est transformée en densité de saillance DS^T via la convolution avec un filtre gaussien.

Enfin, comme précédemment, la fusion des densités de saillance spatiale DS^{SP} et temporelle DS^T fait intervenir un mécanisme de compétition intra et inter cartes, conduisant à l'obtention de la densité de saillance finale DS .

3 Expérimentations oculométriques

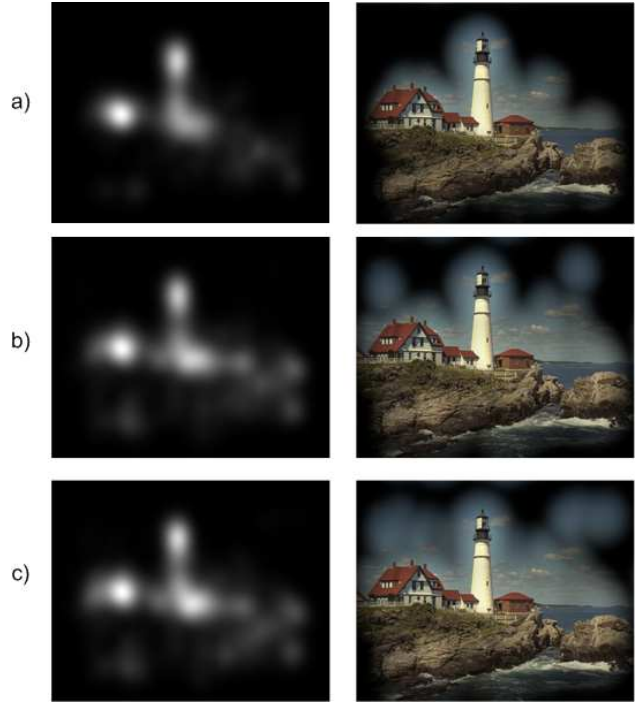


Figure 4 – Densités de saillance humaine a), b) et c) obtenues respectivement pour les temps d'observation 2, 8 et 14 secondes.

Un dispositif oculométrique est un outil permettant de suivre les déplacements de la pupille. La stratégie visuelle d'un observateur est alors aisément interprétable. Ce type de dispositif est basé sur la capacité de nos yeux à réfléchir les infrarouges. En fait, deux types de reflets sont observés : des reflets fixes dus à la réflexion des infrarouges sur la cornée (reflets de Purkinje), et des reflets mobiles dus à la réflexion des infrarouges sur la pupille. La position relative de ces deux types de reflets permet de déterminer le positionnement de l'oeil. Les données recueillies peuvent être exploitées de différentes façons. Les zones attirant le regard des observateurs sont déterminées. La durée de fixation est un élément intéressant pour mesurer le degré de saillance d'une zone. Par ailleurs, la stratégie visuelle, c'est à dire le déplacement oculaire, peut faire l'objet d'étude, même si

Tableau 1 – *Protocole des expérimentations pour l'acquisition de données oculométriques sur images fixes.*

Protocole	
Distance d'observation	$4H$ H hauteur de l'écran
Résolution de l'écran	800×600
Nombre d'images traitées	40
Type d'images	niveaux de gris et couleur
Nombre d'observateurs	40
Durée de l'observation	15s
Calibrage	20 points de calibrage

cela semble a priori difficile à aborder du fait de l'idiosyncrasie de la stratégie visuelle.

Ces expérimentations nous ont permis de construire une référence, que nous appelons également vérité terrain, permettant d'évaluer la pertinence de nos résultats. Les tableaux ?? et ?? donnent respectivement le protocole expérimentale pour les tests oculométriques sur images fixes et sur séquences d'images animées.

Tableau 2 – *Protocole des expérimentations pour l'acquisition de données oculométrique sur séquences d'images.*

Protocole	
Distance d'observation	$5H$ H hauteur de l'écran
Résolution de l'écran	800×600
Nombre de séquences	8
Nombre d'observateurs	30
Calibrage	12 points de calibrage

L'obtention de la séquence de carte de fixation pour un observateur se fait via la relation suivante :

$$CS(x, y; t) = \sum_{i=1}^M \Delta(x - x_i, y - y_i; t - t_i) \quad (10)$$

avec, M le nombre total de fixations pour la durée du test, $(x_i, y_i; t_i)$ les coordonnées de la fixation i et son instant temporelle t_i et Δ le symbole de Kronecker.

A partir des séquences de fixation obtenues pour chaque observateur, le comportement oculomoteur d'un observateur moyen est déterminé en accumulant chaque carte de fixation issue du même instant temporel. L'application d'une gaussienne bi-dimensionnelle permet de construire une densité et de prendre en compte la précision limitée de l'oculomètre. La relation suivante définit cette séquence :

$$DS(x, y; t) = \left(\frac{1}{N} \sum_{i=1}^N CS^i(x, y; t) \right) * g_{\sigma_x, \sigma_y}(x, y) \quad (11)$$

avec CS^i la séquence de fixation de l'observateur i , N le nombre d'observateurs, et g_{σ_x, σ_y} une gaussienne bi-dimensionnelle.

4 Evaluation du modèle

A partir de la vérité terrain, une étude comparative quantitative est menée. Les résultats de l'évaluation du modèle proposée pour les images fixes sont tout d'abord rappelés [?, ?, ?]. L'évaluation des performances du modèle sur un contenu dynamique est ensuite abordée.

4.1 Evaluation de la modélisation spatiale

L'évaluation quantitative de la pertinence des cartes de saillance prédite se fait via le coefficient de corrélation linéaire. Le tableau ?? présente le coefficient de corrélation linéaire calculé sur une base de 18 images entre les cartes de saillances spatiales prédites et la vérité terrain pour différents temps d'observations. Les résultats du modèle de L. Itti [?] sont également donnés. Ces derniers permettent d'apprécier le gain en corrélation de l'approche proposée. La valeur du coefficient de corrélation augmente avec le

Tableau 3 – *Coefficient de corrélation linéaire moyen déterminé sur une base de 18 cartes de saillance obtenues avec le modèle d'attention visuelle spatiale . Les valeurs de corrélation sont données pour différents temps d'observation.*

cc	4s	10s	14s
Approche proposée	0.42	0.47	0.54
L. Itti	0.32	0.35	0.37

temps d'observation ce qui est normal pour deux raisons : tout d'abord, la dépendance temporelle n'est pas prise en compte dans l'actuelle modélisation. Quel que soit le temps d'observation utilisé, le modèle se comporte comme si il disposait d'un temps d'observation infini. Par ailleurs, plus le temps d'observation augmente et plus les observateurs découvrent l'image; le degré de similarité des données réelles et prédites augmente alors. La différence de performance entre le modèle proposé et celui de L. Itti tend à démontrer que la modélisation proposée est biologiquement plausible et efficace.

4.2 Evaluation de la modélisation spatio-temporelle

L'évaluation de la modélisation spatio-temporelle est plus délicate. L'utilisation d'un coefficient de corrélation n'est plus adaptée : chaque image de la séquence est vue par un observateur pendant une très courte durée (20 ms) alors que, comme nous l'avons souligné précédemment, le modèle proposé est indépendant du temps d'observation. Pour remédier à ce problème de dépendance temporelle, nous utilisons un algorithme de classification à deux états, saillant et non saillant, et à population non constante par catégorie. La population des pixels saillants doit être au plus de 30%.

L'utilisation d'une matrice de confusion permet ensuite de déterminer la précision du système, notée AC . Le tableau

?? donne la précision moyenne de la classification pour trois séquences. La valeur moyenne *AC* est calculée sur les 90 premières images de chaque séquence. La précision de la classification est correcte. Plus de 75% des pixels sont bien classés. Ce résultat moyen de 77% est obtenu uni-

	Kayak	Table	Stefan	Moyenne
AC	0,8	0,78	0,75	0,77

Tableau 4 – Précision *AC* moyenne de la classification pour différentes séquences calculée sur 90 images.

quement à partir des caractéristiques visuelles de bas niveaux. Sachant qu’aucune information cognitive n’est prise en compte, ce résultat est tout à fait intéressant.

5 Conclusion

Les performances de nombreuses applications appartenant au domaine du traitement d’images sont susceptibles d’être améliorées via des informations a priori : le rendement d’un codeur vidéo, par exemple, est nettement amélioré lorsque des événements temporels tels que des changements de plan ou des fondus enchaînés sont identifiés ; une configuration particulière du codeur est alors mise en place. Aujourd’hui, la détection de ces événements est bien maîtrisée. Il n’en était pas de même pour les zones dites d’intérêt. La connaissance de la position spatiale de ces zones, très sensibles à tous types de dégradations, va s’avérer primordiale pour la définition de futurs outils du traitement d’images.

Dans ce cadre, nous proposons une modélisation de l’attention visuelle dite exogène opérant sur des séquences d’images. Le résultat de cette modélisation se traduit par l’obtention d’une séquence de cartes de fixation ou l’intérêt de chaque pixel est quantifié. Les résultats sont très encourageants : la modélisation spatiale proposée est meilleure, en terme de corrélation, que celle proposée par L. Itti considérée aujourd’hui comme une référence. Enfin, la dimension temporelle a été ajoutée ce qui rend le modèle encore plus attractif. La précision moyenne du modèle, dans un cadre de classification supervisée, atteint 77%. Compte tenu que des informations de haut niveaux, telles que les visages, le texte, ne sont pas intégrées dans le modèle, ces résultats sont très encourageants.

A moyen terme, des applications telles que le codage vidéo, le tatouage numérique et l’estimation de qualité devraient bénéficier de la connaissance a priori des positions spatiales des régions d’intérêt.