Modèle statistique et description locale d'apparence pour la détection des contours des lèvres.

P. GACON[1], P.-Y. COULON[1], G. BAILLY[2]
[1]LIS, Laboratoire des Images et des Signaux, [2]ICP, Institut de la Communication Parlée, 47 Avenue Félix Viallet, 38031 Grenoble
pierre.gacon@lis.inpg.fr , pierre-yves.coulon@lis.inpg.fr , bailly@icp.inpg.fr

Résumé – La segmentation des lèvres est une tâche importante qui a de nombreuses applications dans divers domaines de recherches tels que la reconnaissance de la parole ou la synthèse de visage réaliste. Dans ce travail, nous portons une intention particulière sur la détection du contour intérieur des lèvres, opération difficile à effectuer avec robestesse du fait de non linéarités. Nous proposons une méthode basée sur un modèle actif de la forme et de l'apparence des lèvres et mettant en oeuvre une prédiction non-linéaire de descripteurs locaux d'apparence. Les caractéristiques d'apparence sont également séparés en une composante statique et dynamique afin d'optimiser le traitement d'un séquence d'image en adaptant le modèle au locuteur.

Abstract – Lips segmentation is an important task with many applications in various research area such as speech recognition or realistic face synthesis. In this work we have a special focus on the detection of the inner mouth contour which is quite difficult to achieve with robustness because of nonlinearities . We propose a method based on an active model for shape and appearance with a nonlinear prediction of local appearance descriptors. The appearance is divided in two components, static and dynamic, in order to adapt the model to the speaker in a video sequence.

1. Introduction

La segmentation des lèvres est une tâche qui trouve des applications dans divers domaines de recherches tel la reconnaissance automatique de la parole (dans le cadre d'interface homme/machine par exemple), l'identification de personnes ou l'augmentation de l'intelligibilité de la parole dans des situations bruitées, les télécommunications à très bas débit ou l'animation d'un avatar d'une personne dans un environnement virtuel.

Si les applications sont nombreuses, cette tâche n'en demeure pas moins ardue. La variabilité des formes de lèvre ou de teinte de peau d'un locuteur à un autre, les changements d'éclairage, les mouvements de lèvres peu habituels ou extrêmes, sont autant d'éceuils qui réduisent la précision et la robustesse. Les méthodologies utilisées sont ainsi aussi nombreuses que variées mais l'on peut tout de même discerner trois grandes catégories : les méthodes sans modèle global, les méthodes avec modèle paramétrique et les méthodes avec modèle statistique, cette dernière approche ayant été souvent favorisée ces dernières années.

Dans le cadre des méthodes sans modèle, Zhang [1] avait par exemple proposé d'utiliser la teinte et une détection de contour pour segmenter la bouche mais sans imposer la moindre contrainte de forme, ce qui conduisait à des résultats grossiers. Delmas [2] utilisa les mêmes informations mais avec une optique de contours actifs (ou snakes) afin d'imposer une certaine régularité à la forme. Cette famille de méthode peut donner des résultats parfaitement satisfaisants si les conditions sont bonnes : éclairage contrôlé, bon contraste entre la couleur de la peau et celle des lèvres. Néanmoins, si les conditions sont plus difficiles, la détection devient moins précise et souffre donc d'un manque de robustesse qui en rend l'utilisation difficile pour des applications ou les contours doivent être connus avec finesse. Surtout, rien n'assure que la forme des lèvres détectées sera cohérente

Dans l'objectif d'avoir un contour détecté plus réaliste, plusieurs auteurs ont proposé d'utiliser des modèles paramètriques. La difficulté de ces modèles est de trouver le bon dosage de flexibilité : un modèle peu flexible donnera toujours une forme de lèvre plausible mais échouera à segmenter des formes de bouches inhabituelles alors qu'un modèle trop flexible donnera des formes non-réalistes dans certains cas.

A titre d'exemple, Hennecke et al. [3] ont introduit les patrons déformables (deformable template). Le patron est un modèle de lèvres contrôlé par un jeu de paramètres dont les valeurs sont choisies en minimisant une fonction de coût. Dans un souci d'avoir un modèle suffisamment flexible afin de pouvoir modéliser n'importe quelle configuration de lèvre, Eveno [4] a proposé d'utiliser des courbes paramétriques cubiques pour décrire les contours en imposant conditions et limites aux dérivées des courbes au niveaux des points saillants.

Au cours des dernières années, l'approche par modélisation statistique des lèvres a été de plus en plus plébiscitée. Cootes et al. [5] ont ainsi proposé d'utiliser les modèles de formes actifs (active shape model). La forme d'un objet est apprise sur une base de donnée contenant de nombreux exemples: on procède alors à une analyse en composantes principales (ACP) de la forme afin de diminuer la taille de l'espace de représentation et de commander le modèle en réglant un nombre limité de paramètres. L'intérêt principal est que l'on générera toujours une forme plausible, la seule limitation à la flexibilité du modèle étant la taille de la base d'apprentissage (une configuration absente des exemples pouvant être mal modélisée). Par la suite Cootes et al. [7] ont proposé d'étendre le principe à la description de l'apparence avec les modèles d'apparence actifs (active appearance model) dans lesquelles non seulement la forme de la bouche mais aussi les valeurs des niveaux de gris de la région d'intérêt sont modélisées.

Dans notre travail, cette approche a été retenue et nous avons plus particulièrement dirigé nos recherches sur deux axes principaux. Le premier est la précision de la détection du contour intérieur des lèvres qui présente un comportement non linéaire (ouverture de la bouche, apparition/disparition des dents) et qui est donc moins bien modélisé par un modèle fondé sur une analyse en composantes principales que le contour extérieur. Le second point est une volonté de personnaliser

au fil d'une séquence d'image le modèle au locuteur traité, afin d'augmenter la précision et la rapidité de convergence du modèle.

Pour ce faire nous avons séparé l'apparence globale en deux quantités: une apparence "statique" qui est caractéristique d'un locuteur donné (et qui sera initialisée sur la première image d'une séquence grâce à une classification sommaire des pixels) et une apparence "dynamique" qui correspond aux fluctuations induites par le mouvement (et donc plus particulièrement la parole). En outre, les critères classiques de maximisation du flux du gradient à travers une courbe s'appliquant mal aux contours intérieurs de la bouche, on utilisera des descripteurs locaux gaussiens dont les réponses seront prédites non-linéairement.

2. Base d'apprentissage et données

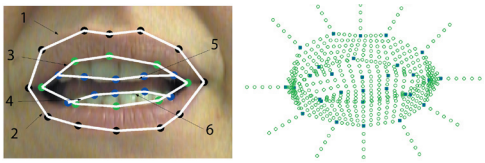

Figure 1 : Exemple d'image annotée de la base d'apprentissage et grille utilisée pour l'apprentissage de l'apparence.

La base d'apprentissage est constituée de séquences vidéo de 12 locuteurs différents. Afin de construire le modèle, N=450 images ont été annotées manuellement en 30 points de contrôle décrivant la forme: 12 points sur le contour externe des lèvres et 8 points sur le contour interne ainsi que 10 points sur les dents (lorsque les dents ne sont pas visibles, les points correspondant existent toujours mais sont confondus avec le contour intérieur de la bouche). Les coordonnés des points pour chaque image sont sauvegardées dans les vecteurs de forme $\boldsymbol{s}_i$ ($1 \le i \le N$). L'opérateur assigne également à chaque image un Etat Général de la Bouche (EGB). L'EGB est une variable d'état décrivant sommairement 4 états de la bouche (fermée, ouverte, grande ouverte, sourire).

Comme les notions de couleur et de luminosité sont mélangées dans l'espace couleur RVB, nous utilisons l'espace couleur YCbCr où la chrominance et la luminance sont décorellées. L'apparence échantillonnée (c'est à dire les 3 valeurs YCbCr associées à un pixel) est extraite de l'image en 728 points d'une grille d'apprentissage déduite des 30 points de contrôle (cf figure 1). L'utilisation de la grille définit précisément si un pixel correspond à de la peau, aux lèvres, aux dent ou au fond de la bouche. Cela est particulièrement utile pour décrire l'intérieur de la bouche et éviter un effet de flou.

Dans notre méthode, nous faisons la distinction entre l'apparence statique et l'apparence dynamique. La première correspond aux caractéristiques spécifiques d'un locuteur donné (couleurs des lèvres et de la peau et contraste entre les deux) et donc à la variabilité extrinsèque de l'apparence. On la définit comme l'apparence moyenne d'un locuteur donné et elle est stockée dans les vecteurs $\boldsymbol{a}_{m,i}$ ($1 \le i \le N$). A l'inverse, l'apparence dynamique correspond aux variations intrinsèques et est la différence entre l'apparence totale et l'apparence statique. Elle est stockée dans les vecteurs $\boldsymbol{a}_{d,i}$ ($1 \le i \le N$).

Enfin, notons que notre modèle n'étant qu'en deux dimensions il ne tolère donc que de petites rotations du visage.

3. Modélisation

3.1 Modèle actif de forme et d'apparence

On procède ici à l'analyse en composante principale des données collectées dans la partie 2. Par exemple dans le cas de la forme (contenue dans les vecteurs $\mathbf{s}_i$) on calcule le vecteur moyen $\overline{\mathbf{s}}$ puis la matrice de covariance $\mathbf{S}$. Les vecteurs propres de $\mathbf{S}$ correspondent aux différents modes propres de variations, ceux associés à de grandes valeurs propres modélisant une part plus importante de la variance des données.

$$\bar{\boldsymbol{s}} = \frac{1}{\mathrm{N}}\sum_{i=1}^{\mathrm{N}} \boldsymbol{s}_i \;;\; \mathbf{S} = \frac{1}{\mathrm{N}}\sum_{i=1}^{\mathrm{N}} \left(\boldsymbol{s}_i - \bar{\boldsymbol{s}}\right)\left(\boldsymbol{s}_i - \bar{\boldsymbol{s}}\right)^T$$

On choisit de garder 95% de la variance totale, les vecteurs propres correspondant étant sauvés dans la matrice $\mathbf{P}_S$. Finalement, on peut générer n'importe quelle forme s en ajustant simplement le vecteur $\boldsymbol{\sigma}$ (contenant les poids affectés à chaque mode propre) dans l'équation suivante: $\boldsymbol{s} = \overline{\mathbf{s}} + \mathbf{P}_s\boldsymbol{\sigma}$. On procède de même avec les deux apparences puis on procède à une ACP de deuxième niveau similairement à [6] afin d'avoir un modèle combiné modélisant de façon cohérente la forme et l'apparence dynamique. On obtient le jeu d'équations suivant:

$$(1)\ \mathbf{c} = \begin{bmatrix} \mathrm{W}.\boldsymbol{\sigma} \\ \alpha_d \end{bmatrix} = \bar{\boldsymbol{c}} + \mathbf{P}_c\chi \;\Rightarrow\; \begin{cases} (2)\ \boldsymbol{s} = \bar{\boldsymbol{s}} + \mathbf{P}_s\boldsymbol{\sigma} \\ (3)\ \boldsymbol{a}_d = \overline{\boldsymbol{a}_d} + \mathbf{P}_d\alpha_d \end{cases}$$

$$(4)\ \boldsymbol{a}_m = \overline{\boldsymbol{a}_m} + \mathbf{P}_m\alpha_m$$

L'équation (2) contrôle donc le modèle de forme, l'équation (3) contrôle l'apparence dynamique et l'équation (4) l'apparence statique. Enfin l'équation (1) correspond au modèle combiné forme/apparence dynamique avec W un coefficient normalisant les unités. Segmenter les lèvres sur une image revient finalement à optimiser le jeu de paramètres correspondant aux modes principaux des ACP retenus, soit 18 dans notre cas : 9 dans le vecteur χ et 9 dans le vecteur $\boldsymbol{\alpha}_\mathbf{m}$.

Les valeurs moyennes de χ sont également calculées pour chaque EGB et seront utilisées par la suite.

3.2 Descripteurs locaux, principe et validation

La maximisation d'un flux de vecteur gradient à travers les courbes des lèvres est un critère souvent retenu pour en accomplir la segmentation. Néanmoins il s'applique mal aux contours intérieurs du fait du comportement non linéaire de cette zone (fermeture/ouverture, apparition du contour des dents ce qui peut entre autre entraîner la disparition ou l'inversion de la direction des vecteurs gradients).

Pour résoudre ce problème, nous avons utilisé des descripteurs locaux d'apparence basés sur les réponses de filtres dérivés gaussiens (limités à la première dérivée).

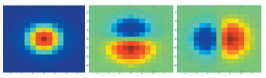

Figure 2 : Réponses impulsionelle des filtres gaussiens G (moyenne), Gx et Gy (gradients horizontal et vertical)

Le critère calculé est la différence entre la réponse de ces descripteurs le long des contours et une prédiction de leur réponse faite pour une configuration donnée de la forme.

Cette prédiction est effectuée par un réseau de neurones non linéaire dont l'entrée est la forme et dont la dimension de l'espace de sortie (la réponse des filtres) est diminuée par ACP afin de réduire la taille du réseau et les temps de calcul. Ainsi

forme et apparence sont décrites par des modèles linéaires mais leur couplage est non-linéaire.

Nous avons testé validé la pertinence de cette idée en comparant ses performances de segmentation à celles d'autres méthodes proches. Ces comparaisons se feront dans un cas mono-locuteur (avec une base d'apprentissage de 50 images). La variabilité inter-locuteur n'est donc pas prise en compte, l'apparence statique correspondant simplement à l'apparence moyenne du locuteur. La réponse des filtres ou les valeurs des pixels dépendent ainsi uniquement du mouvement des lèvres et des différences d'illumination, effet qui peut être atténué en filtrant la luminance par le filtre rétine [7] (opérateur permettant de diminuer les variations d'illuminations sur une image en utilisant un modèle s'inspirant de l'oeil humain).

Dans tous les modèles, la forme sera décrite par un modèle actif obtenu par une ACP. En revanche l'apparence sera décrite par A) les descripteurs gaussiens locaux ou B) les valeurs des pixels obtenus grâce à la grille d'échantillonnage. Pour chacune de ces modélisations de l'apparence, nous allons tester trois types de méthode: 1) un modèle d'apparence classique type [6] (ACP sur l'apparence puis ACP de second niveau pour avoir un modèle combiné forme/apparence), 2) l'apparence est prédite à partir des paramètres du modèle de forme grâce à une régression linéaire (entraînée sur la base d'apprentissage), 3) l'apparence est prédite à partir des paramètres du modèle grâce à un réseau de neurones non-linéaire. Il est à noter que pour les méthodes 2) et 3) le nombre de paramètres à optimiser est inférieur à la méthode 1) puisque seule la forme est optimisée.

Nous testons ensuite les 6 cas sur 50 images non présentes dans la base d'apprentissage. Les erreurs sont données en pourcentage de la largeur de la bouche.

méthode	contour extérieur	contour intérieur	dents	tous points	nombre d'itérations
A) 1)	1.6 / 1.1	2.4 / 1.7	2.6 / 1.8	2.1 / 1.5	29.8
2)	1.7 / 1.2	2.9 / 1.8	3.1 / 2.2	2.4 / 1.7	17.2
3)	1.4 / 0.8	1.8 / 0.9	1.8 / 0.9	1.5 / 0.8	9.8
B) 1)	1.4 / 0.9	2 / 1.1	2.2 / 1.2	1.8 / 1	31.1
2)	1.5 / 1	2.2 / 1.3	2.5 / 1.4	2 / 1.2	15.1
3)	1.5 / 0.9	2.1 / 1.3	2.4 / 1.3	2 / 1.1	15.5

Tableau 1: Résultats donnés en pourcentage: erreur/ écart type. Méthodes: A) apparence décrite par descripteurs locaux, B) apparence décrite par pixels, 1) ACP sur l'apparence, 2) apparence prédite par régression linéaire sur la forme, 3) apparence prédite par réseau de neurones.

Les résultats montrent que les meilleurs résultats sont obtenus pour la méthode proposée dans notre contribution.

On constate que les 3 méthodes où l'apparence est décrite par la grille d'échantillonnage donnent des résultats très proches les unes des autres, l'utilisation de la grille permettant de réduire les problèmes dus aux non-linéaritées grâce à sa description précise de la zone intérieure de la bouche. La méthode B)1) donne alors les meilleurs résultats, au prix d'un nombre supérieur d'itérations dû au plus grand nombre de paramètres à régler.

L'utilisation des descripteurs locaux ne donne de résultats véritablement intéressants que dans le cas de la prédiction non-linéaire. Dans les cas 1) et 2), la linéarité de la relation forme/ apparence résulte en effet en une mauvaise segmentation de l'intérieur de la bouche, ce qui entraîne également des erreurs sur le contour extérieur. En revanche, avec la prédiction non linéaire, les descripteurs locaux permettent d'obtenir de meilleurs résultats du fait qu'ils sont moins sensibles au bruit qu'une mesure directement liée aux pixels et que leur information de type gradient leur permet de coller au plus près des contours.

La validation ayant été probante, ce principe a été étendu au cas multi-locuteur. Pour tenir compte de la variabilité extrinsèque rajoutée, l'apparence statique (qui est justement caractéristique de cette variabilité) a alors été rajoutée en entrée du réseau de neurone.

Pour effectuer la prédiction nous avons décidé d'utiliser un réseau à deux couches construit par rétropropagation. Nous avons donc les 18 entrées correspondant aux paramètres du modèle, 15 unités cachées et 15 sorties correspondant à l'espace réduit obtenu grâce à l'ACP sur la réponse des filtres.

3.3 Classification initiale par modèle couleur

Un modèle colorimétrique est utilisé pour discriminer grossièrement les pixels de lèvres et de peau, ceci afin d'initialiser l'apparence statique sur la première image d'une séquence d'un locuteur donné. A partir de la base d'apprentissage annotée, les valeurs CbCr de chacune des deux classes peau et lèvre sont extraites puis leur distribution statistique est modélisée par une mixture de gaussienne. A partir d'un simple calcul de probabilité, on peut alors déterminer si un pixel a plus de chance d'être de la peau ou des lèvres (on fixe également un seuil assez restrictif sur la valeur de la probabilité pour ne pas assigner les pixels incertains). On obtient alors une segmentation grossière mais suffisante pour l'utilisation qui en est faite.

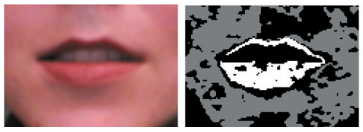

Figure 3 : Exemple de classification de pixels (blanc: lèvre, gris: peau, noir: non affecté)

3.4 Modèle de commissures des lèvres

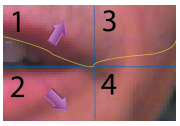

Figure 4 : Zone des commissures des lèvres et ligne des minima de luminance (en jaune).

Les commissures des lèvres sont considérées comme des points-clés dont la connaissance détermine la position et l'échelle de la bouche. Ces points sont délicats à détecter car ne se trouvant généralement pas sur des contours nets, mais au contraire dans des zones floues ou ombrées. Ainsi, ils seront vus comme les points de jonction de quatre quadrants de caractéristiques différentes (repérés sur la figure 4). Les régions 1 et 2 sont non homogènes car présentant un contour entre les lèvres et la peau au contraire des régions 3 et 4. Chacune de ces régions sera décrite par un jeu de filtres dérivés gaussiens dont les réponses seront modélisées statistiquement par une mixture de gaussiennes.

Si l'on peut considérer, comme cela a déjà été proposé par [4], que les points de commissures se trouvent sur le ligne des minima de luminance par colonne (la ligne jaune sur la figure

4), il reste seulement à trouver l'indice des colonnes. On teste alors les points de la ligne d'intérêt et l'on sélectionne finalement le couple de points le plus probable comme commissures des lèvres. Des valeurs initiales pour les paramètres S et T, correspondant à l'échelle et la position des lèvres, sont ainsi obtenus. La méthode est robuste dans plus de 95% des cas pour une précision correspondant à 5% de la taille réelle des lèvres.

4. Segmentation

Lors du traitement d'une image inconnue, nous voulons segmenter les lèvres en trouvant le meilleur jeu de paramètres χ, α_m, S et T pour notre modèle actif. Cette tâche est effectuée en minimisant une fonction de coût grâce à l'algorithme classique du simplex. Cette fonction est la somme pondérée de deux erreurs quadratiques moyenne (EQM): l'EQM entre la réponse observée sur l'image des descripteurs locaux et leur réponse prédite et l'EQM entre l'apparence modélisée et l'apparence observée sur l'image. Les paramètres sont considérés avoir convergé dès que la différence entre les valeurs maximum et minimum du simplex passe en dessous d'un seuil.

Pour la première image d'une séquence, les paramètres du simplex sont initialisés en testant les jeux de paramètres moyens correspondant aux différents EGB (pour χ), en utilisant la classification de pixel (pour α_m) et en utilisant le modèle de commissures (pour S et T). Par la suite, les valeurs finales des paramètres des images précédentes sont utilisées pour obtenir les valeurs initiales.

Au fil de la séquence, α_m va converger (l'intervalle de recherche des paramètres étant diminué progressivement) ce qui correspond au fait que les caractéristiques du locuteur sont de mieux en mieux connues. Lorsque la convergence a eu lieu (la variation d'α_m passe en-dessous d'un seuil), la fonction de coût sera calculée comme étant uniquement l'EQM sur les descripteurs, seul le vecteur de paramètre χ sera dès lors optimisé à chaque nouvelle image. Le principe est résumé par la figure 5.

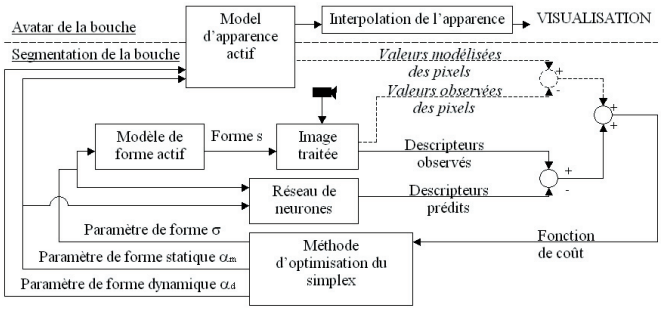

Figure 5 : Principe de la segmentation.
En pointillé: calcul effectué tant que α_m n'a pas convergé.

5. Résultats

La table 1 présente quelques résultats de segmentation chiffrés, tandis que les figures 6 et 7 montrent des exemples de segmentation et de génération d'avatar par interpolation de l'apparence échantillonnée. Au final, notre méthode produit des résultats fiables et robustes et semble particulièrement adaptée pour modéliser avec fidélité les contours intérieurs de la bouche.

Position de l'erreur	Contour extérieur	Contour intérieur	Dents	Total
Test sur la base d'apprentissage	2.7 / 1.3	2.8 / 1.4	3 / 1.4	2.8 / 1.3
Locuteur dans la base, image inconnue	2.7 / 1.3	2.8 / 1.5	3.1 / 1.4	2.8 / 1.3
Leave-one-out	3 / 1.4	3.1 / 1.4	3.3 / 1.5	3.1 / 1.3

Tableau 2 :
Résultats donnés en pourcentage: erreur/écart type, leave-one-out: locuteur testé enlevé de l'apprentissage

Figure 6 : Exemples de segmentation

Figure 7 : Exemples d'interpolation d'avatars

Une évaluation subjective de la méthode est en cours d'expérimentation en quantifiant l'intéllégibilité du locuteur synthétisé (prononçant des numéros de téléphone pour divers niveaux de bruit sur le canal audio).

6. Références

[1] X. Zhang, R. M. Mersereau, M. A. Clements and C. C. Broun, "Visual Speech Feature Extraction for Improved Speech Recognition", In Proc. ICASSP'02, 2002, pp. 1993-1996

[2] P. Delmas, N.Eveno, and M. Lievin, "Towards Robust Lip Tracking", *International Conference on Pattern Recognition (ICPR'02),* Québec City, Canada, Août 2002

[3] M. Hennecke, V. Prasad, and D. Stork. "Using deformable templates to infer visual speech dynamics", *28 h Annual Asimolar Conference on Signals, Systems, and Computer*, *volume 2*, IEEE Computer, Pacific Grove, pages 576-582, 1994.

[4] N.Eveno, A. Caplier, and P-Y Coulon, "Automatic and Accurate Lip Tracking", *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 14, no.5, pp. 706-715, Mai 2004

[5] T.F. Cootes, C.J. Taylor, D.H. Cooper, "Active Shpae Models - Their Training and Application", Computer Vision and Image Understanding, Vol. 61, No. 1, Janvier, pp. 38-59, 1995

[6] T. F. Cootes. "Statistical models of appearance for computer vision", Rapport technique disponible en ligne sur http://www.isbe.man.ac.uk/bim/refs.html, 2001.

[9] W.H.A. Beaudot, "The neural information processing in the vertebrate retina: A melting pot of ideas for artificial vision", PhD Thesis in Computer Science, INPG (France) décembre 1994