UNIVERSITÉ DE MACÉDOINE

DÉPARTEMENT D'INFORMATIQUE APPLIQUÉE

Thèse de Doctorat de

MANITSARIS Sotirios

Vision par ordinateur pour la reconnaissance des gestes : analyse et modélisation stochastique du geste dans l'interaction musicale

synopsis étendu

Composition du Jury :

G. PEKOS	Directeur de la thèse (Professeur)
A. KATOS	Rapporteur (Professeur)
G. STEPHANIDES	Rapporteur (Professeur Agrégé)
K. KARANIKAS	Examinateur (Professeur)
A. CHATZIGEORGIOU	Examinateur (Professeur Adjoint)
L. STAMOU	Examinateur (Professeur Adjoint)
C. GEORGIADIS	Examinateur (Professeur Adjoint)

Thèse préparée au sein du Laboratoire des Technologies Multimédia et de l'Infographie

Thessalonique 2010

Grèce

Vision par ordinateur pour la reconnaissance des gestes : analyse et modélisation stochastique du geste dans l'interaction musicale.

Résumé

Cette thèse présente un système prototype de vision par ordinateur pour la reconnaissance des gestes dans l'interaction entre le pianiste et l'instrument. La vision par ordinateur est la seule technologie permettant la reconnaissance des gestes, sans interférence entre le pianiste et son instrument, et à un faible coût. Le système propose deux approches pour la reconnaissance : a) l'approche statique, ou reconnaissance des doigtés, et b) l'approche dynamique, extension de l'approche statique. La reconnaissance statique s'applique à chaque image de la vidéo. Elle repose sur l'analyse et l'interprétation des caractéristiques de l'image, en les comparant avec le modèle déterministe du geste. La reconnaissance dynamique s'applique à un ensemble de séquences d'images vidéo. Elle se base sur l'analyse et la modélisation stochastique du geste, à l'aide de Modèles de Markov Cachés. Cette méthode peut être étendue à d'autres champs d'application tels que la surveillance de personnes en perte d'autonomie à domicile, la valorisation du patrimoine culturel, l'étude du comportement humain ou encore l'interaction homme-machine.

Mots-Clés: Vision par ordinateur, modélisation, reconnaissance, geste, signal vidéo, Modèles de Markov Cachés, interaction musicale.

Sommaire

1 Introduction

L'interprétation musicale est le fruit de la symbiose entre le musicien et son instrument musical. Cette symbiose prend la forme d'une relation interactionnelle et gravitationnelle, laquelle peut être symbolisée par un triangle dont la perception, la connaissance et le geste constituent les sommets. Le musicien est à la fois un élément déclencheur et émetteur, reliant les trois sommets du triangle au travers de plusieurs mécanismes communicationnels. Le travail élaboré dans cette thèse de doctorat contribue à l'analyse et la compréhension des mécanismes combinant le geste cinétique et le geste sonore, voire le geste musical, afin de modéliser et de reconnaître des gestes musicaux dans une séquence d'images.

L'analyse et la reconnaissance des gestes musicaux dans une vidéo permettent la compréhension approfondie de l'expressivité musicale. L'art d'interagir musicalement est pour l'artiste un espace d'expression et d'interprétation libre. Le musicien devient une source intarissable de sentiments, exprimés au travers des gestes [Decroux 1994]. En parallèle, les modèles stochastiques peuvent à la fois décrire et interpréter le geste musical et ses éléments structurels, sans tenir compte des sentiments. La vidéo, en tant que séquence d'images, peut contribuer à révéler la structure du geste [Bérard 2000]. L'information de l'expression musicale doit être extraite par la vidéo, et modélisée à l'aide de méthodes stochastiques. Informatique appliquée et mathématiques peuvent ainsi mettre leurs méthodes et techniques au service de l'expressivité musicale.

2 Objectifs

L'objectif général de la présente thèse est la proposition d'une méthodologie de vision par ordinateur et le développement d'un système prototype PianOrasis (*associant les vocables Piano et Orasis - signifiant « vision » en grec -*) pour la reconnaissance des gestes musicaux des doigts.

La méthodologie suit plusieurs objectifs spécifiques (*critères*).

Elle doit être :

1. **Capable** : capable de calculer tous les paramètres définissant les gestes des doigts ;
2. **Vision orientée** : orientée vers l'image du musicien, sans analyse préliminaire ;

3. **Non intervenante** : le musicien doit se sentir libre, sans exigence d'équipement spécifique ;
4. **Accessible** : elle doit être à faible coût, permettant l'utilisation à grande échelle.

3 Etat de l'art

L'analyse du geste constitue un champ de recherche pour plusieurs domaines scientifiques, tels que la Communication Interpersonnelle (CI), l'Interaction Homme-Machine (IHM) ou l'Interaction Musicale (IM). Le contenu du mot «geste», ainsi que la terminologie utilisée pour sa définition, varie selon le domaine scientifique. Dans le domaine de la CI, le geste tend à contenir plutôt une dimension communicative (*expressions, émotions, etc.*), dans le domaine de l'IHM, il se rapporte plutôt à une manipulation (*doigté*), tandis que dans le domaine de l'IM, il combine à la fois une dimension communicative et la manipulation elle-même.

De nombreux travaux ont été menés sur l'étude du geste musical. Ces travaux se classifient en trois catégories : (a) la mesure du geste; (b) la reconnaissance des gestes et (c) l'extraction des doigtés. Dans le premier cas l'objectif est de calculer des mesures définies décrivant le déplacement de certains membres du corps humain. La reconnaissance du geste du musicien parmi un ensemble de gestes préenregistrés, est l'objectif de cette approche, tandis que dans l'extraction des doigtés les études portent sur l'effet du mouvement du doigt, en tant qu'événement discret dans le temps. De la mesure du geste jusqu'à l'extraction des doigtés, les différentes approches sont de plus en plus ciblées. De nos jours, la notion de «reconnaissance des gestes» est la plus usitée, tout en étant représentative des trois domaines de recherche mentionnés.

3.1 Mesure du geste

Les Systèmes de Capture Optique des Mouvements (SCOM), tels que Vicon Peak ou Optitrack, ont déjà été appliqués dans l'analyse de la marche, la rééducation des handicapés ainsi que dans la réalisation d'effets spéciaux pour le cinéma d'animation en 3D [Vicon Peak 2005]. Palmer (2000) a attaché des marqueurs réflectifs sur le vêtement d'un pianiste afin de capter ses mouvements expressifs par mesure de déplacement des marqueurs [Palmer 2000].

Dans un autre cas, l'Institut de Recherche et de Coordination Acoustique/Musique (IRCAM – Paris), en coopération avec l'Université McGill, a mené une recherche sur la mesure optique de capture des mouvements des violonistes en 3D, en utilisant le système Vicon 460 [Rasamimanana 2008, 2009 ; Demoucron 2008]. L'objectif de cette recherche a été la modélisation de l'interprétation musicale en obtenant des informations sur les mouvements du violoniste. Ces systèmes sont souvent utilisés pour la mise en œuvre d'une analyse du geste dans un temps différé par rapport à celui de l'interprétation musicale. Cette méthodologie se base sur des plateformes commerciales spécifiques, présupposant un matériel cher et restrictif pour le musicien.

3.2 Reconnaissance des gestes

Il apparaît difficile de développer des systèmes de reconnaissance de gestes répondant à tous les besoins des utilisateurs, ces besoins variant selon le type d'application. L'information gestuelle délivrée en temps réel par les capteurs embarqués [Aylward 2006 ; Coduys 2004], comme dans le cas de la manette Wii [Grunberg 2008], est sans doute de très haut niveau. Même si cette technologie est souvent utilisée pour la reconnaissance des gestes effectués dans l'espace, il serait pourtant pratiquement impossible qu'elles soient appliquées dans la reconnaissance des gestes des doigts sur une surface ou un objet, puisque les musiciens se sentiraient extrêmement contraints.

Avec ou sans fil, le coût de la technologie des capteurs embarqués a sensiblement diminué ces dernières années, contribuant ainsi au développement de l'analyse du geste dans les arts du spectacle. Boukir et Chenevière (2004) ont mené des recherches pour la reconnaissance d'un ensemble des gestes dansés de ballet contemporain, basée sur les trajectoires des mouvements fournies par des SCOM. L'IRCAM a développé un Réseau de Capteurs Sans Fil (RCSF) pour le suivi continu et la reconnaissance des gestes dansés et musicaux en temps réel [Bevilacqua 2007]. Dans l'application du «violon augmenté», l'architecture matérielle comprenait des capteurs d'accélération, des gyroscopes et un capteur de pression monté sur l'archet du violon, tandis qu'un bracelet autour du poignet du violoniste intégrait l'alimentation et l'émetteur sans fil ZigBee. Selon cette méthode, deux types d'informations sont mises en évidence de façon continue : (a) la similarité (*vraisemblance*) du geste effectué avec d'autres gestes préenregistrés et (b) la progression temporelle du geste effectué [Bevilacqua 2007, 2010].

3.3 Extraction des doigtés

Les études entreprises autour de l'extraction des doigtés portent sur quatre axes : (a) le prétraitement à l'aide de l'analyse de la partition [Radicioni 2004] ; (b) l'analyse en temps réel basée sur la technologie de MIDI [Verner 1995] ; (c) le post-traitement du signal sonore [Traube 2004] et (d) les méthodes de la VpO [Burns 2006]. Toutes ces approches reconnaissent l'effet du mouvement du doigt en tant qu'événement discret, autrement dit il s'agit d'une reconnaissance statique des gestes des doigts. Elles ne tiennent pas compte de la nature stochastique du geste et elles ne peuvent pas ainsi être appliquées dans les interprétations vivantes.

3.4 Conclusions

L'étude approfondie des trois approches pour l'analyse du geste dans l'interaction musicale nous amène aux conclusions ci dessous :

1. la mesure du geste se base sur des systèmes commerciaux spécifiques et elle est appropriée pour l'analyse en temps différé et non pour les interprétations vivantes ;
2. la reconnaissance des gestes à l'aide des RCSF a un coût réduit mais ne peut pas être appliquée pour les gestes des doigts ;
3. les doigtés peuvent être extraits en utilisant des technologies à faible coût mais l'information délivrée se rapporte à des événements discrets, sans tenir compte de la nature stochastique des gestes ;
4. la VpO est la technologie la plus souple pour mettre en œuvre une méthodologie de captation et de reconnaissance des gestes musicaux des doigts.

4 Méthodologie et PianOrasis

L'architecture matérielle de base du système PianOrasis est constituée d'un ordinateur iMac Intel Core 2 Duo 2,66 Ghz, 4 Go Ram et d'une caméra Logitech Quickcam Communicate Deluxe.

L'objectif principal de la méthodologie est la reconnaissance des gestes musicaux des doigts effectués sur un instrument de musique classique (*piano, clarinette etc*.), ou sur un clavier numérique (*synthétiseur*) ou bien dans l'espace, sans instrument de musique. Pour que PianOrasis puisse reconnaître ce genre des gestes, le synthétiseur CS1x de Yamaha a

été utilisé pour le calcul des seuils d'appui sur une touche et pour l'analyse de l'effet de mauvaise détection de «doigt caché», derrière une touche noire du piano, dans la reconnaissance du geste.

4.1 Détection de la peau

Afin de rendre le système capable de détecter la peau dans une vidéo, un Modèle de la Peau (MP) a été développé (Figure 4.1.1). Par obtention d'échantillons des pixels de couleur de la peau et d'ongles extraits de la Photothèque du Pianiste (PP), la Région d'Intérêt (RI) a été déterminée. La normalisation de la RI, autrement dit la conversion de l'espace RGB vers l'espace normalisé rg, rend PianOrasis moins dépendant des variations de luminosité et permet d'identifier le MP en tant qu'un ensemble de valeurs. Le résultat exporté est une image binaire contenant soit la valeur 1 pour les pixels de peau (*avant-plan*), soit la valeur 0 pour le reste des couleurs

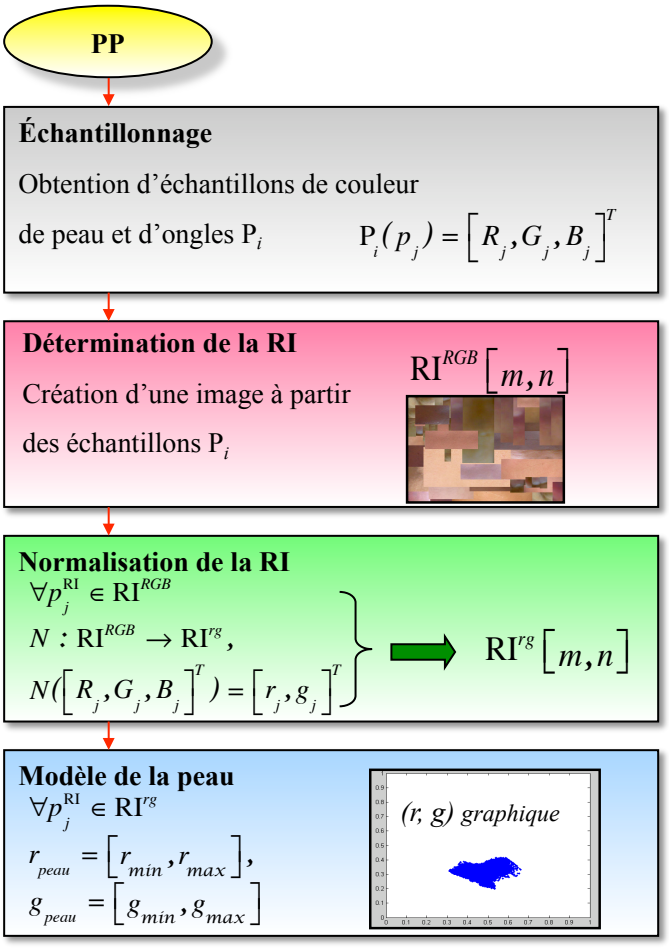

Figure 4.1.1. Création du modèle de la peau

(*arrière-plan*).

Par la suite, une séquence d'images binaires a été créée à partir de la vidéo importée, déterminant ainsi les régions contenant de l'information de peau et d'ongles dans l'image. Parfois, bien que le MP n'étant pas parfait, de petites zones de l'arrière-plan sont considérées par le système comme si elles appartenaient à l'avant-plan et vice versa. Ce problème peut être résolu en appliquant des méthodes de morphologie mathématique pour la réduction du bruit.

4.2 Segmentation de la main

La main du pianiste prend une posture semi-étendue durant son interprétation, augmentant ainsi le niveau de difficulté dans la reconnaissance. Vue de face (*vue de devant pour la caméra*), la zone intérieure de la main étant également une région de peau, dans plusieurs cas la silhouette de la main est extraite en masse avec du bruit. En conséquence, la distinction des doigts dans l'image devient extrêmement difficile, surtout si la distance entre les bouts des doigts est très faible (Figure 4.2.1).

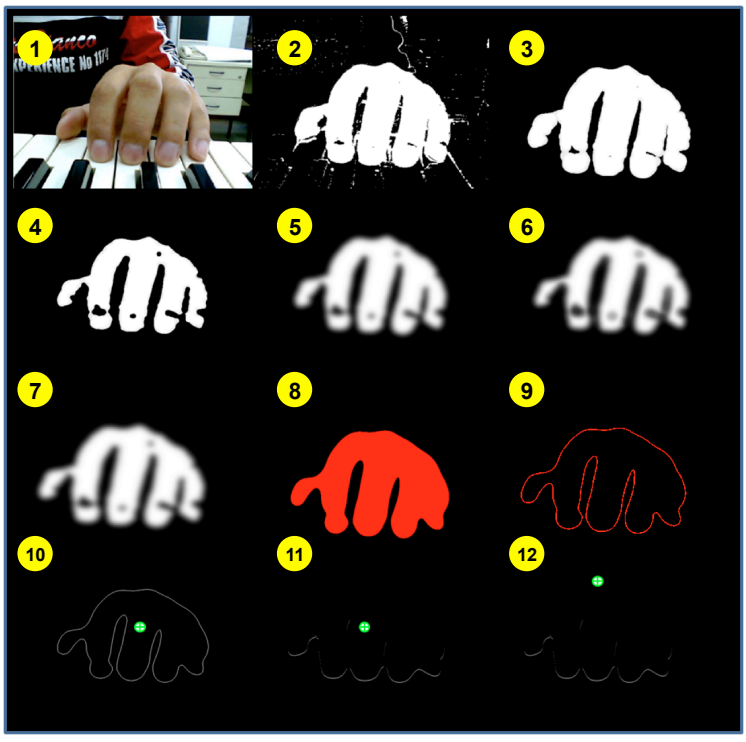

Figure 4.2.1. Segmentation de la main et détection des doigts

(1) image initiale de la vidéo ; (2) application du modèle de la peau ; (3) filtre alternatif séquentiel ; (4) 1ère dilatation; (5) 1er filtre Gauss ; (6) 2ème dilatation; (7) 2ème filtre Gauss ; (8) seuillage ; (9) extraction du contour ; (10) calculation du centroïde ; (11) localisation des bouts des doigts ; (12) mise à zéro de l'ordonnée du centroïde.

Pour cela, l'image binaire est importée dans l'algorithme de la segmentation de la main afin qu'un ensemble de méthodes de traitement d'image lui soit appliqué, comprenant (a) la *simplification de l'image binaire* par réduction de bruit et extraction de la silhouette de la main et (b) la *décomposition de l'image* par extraction du contour de la main et des bouts des doigts [Papamarkos 2000].

4.3 Localisation des doigts

Plusieurs algorithmes de localisation/identification individuelle des doigts dans l'image, utilisant diverses techniques de détection telles que la projection des signatures, la transformée de Hough, les marqueurs colorés, aussi bien que d'autres basées sur des propriétés géométriques, ont été développés.

Le nouvel algorithme, développé dans le cadre de la recherche doctorale et mis en oeuvre dans PianOrasis, suit les critères de détection et de localisation définis par Canny, tout en exploitant les propriétés géométriques de la posture de la main en palmier semi-étendu. Les autres techniques ne sont pas forcément satisfaisantes car elles ne localisent qu'indirectement les bouts des doigts, augmentant ainsi la puissance de calcul nécessaire.

La localisation des doigts s'effectue en calculant les distances Euclidiennes entre le centroïde et les coordonnées des pixels appartenant au contour des doigts. Le calcul des maxima locaux des distances Euclidiennes contribue à l'identification des doigts. Dan le cas d'un «doigt caché», PianOrasis prévoit la position du doigt dans l'image suivante à l'aide des classificateurs, en tenant compte de la «mémoire du geste», calculée en continu par les positions des doigts dans les trois images précédentes (Figure 4.3.1).

4.4 Extrait de vecteurs d'observation

A partir du moment où le centroïde est calculé et les bouts des doigts sont identifiés et localisés dans l'image, PianOrasis peut extraire les vecteurs d'observations, en fonction desquels la reconnaissance des gestes sera effectuée.

Les vecteurs d'observation enregistrés par PianOrasis sont : (a) les différences entre l'ordonné de chaque doigt et celle du centroïde ; (b) les abscisses des doigts et (c) les différences entre les abscisses des doigts adjacents. La reconnaissance statique relie uniquement les vecteurs du pre-

mier cas, tandis que la reconnaissance dynamique tient compte des tous les trois cas.

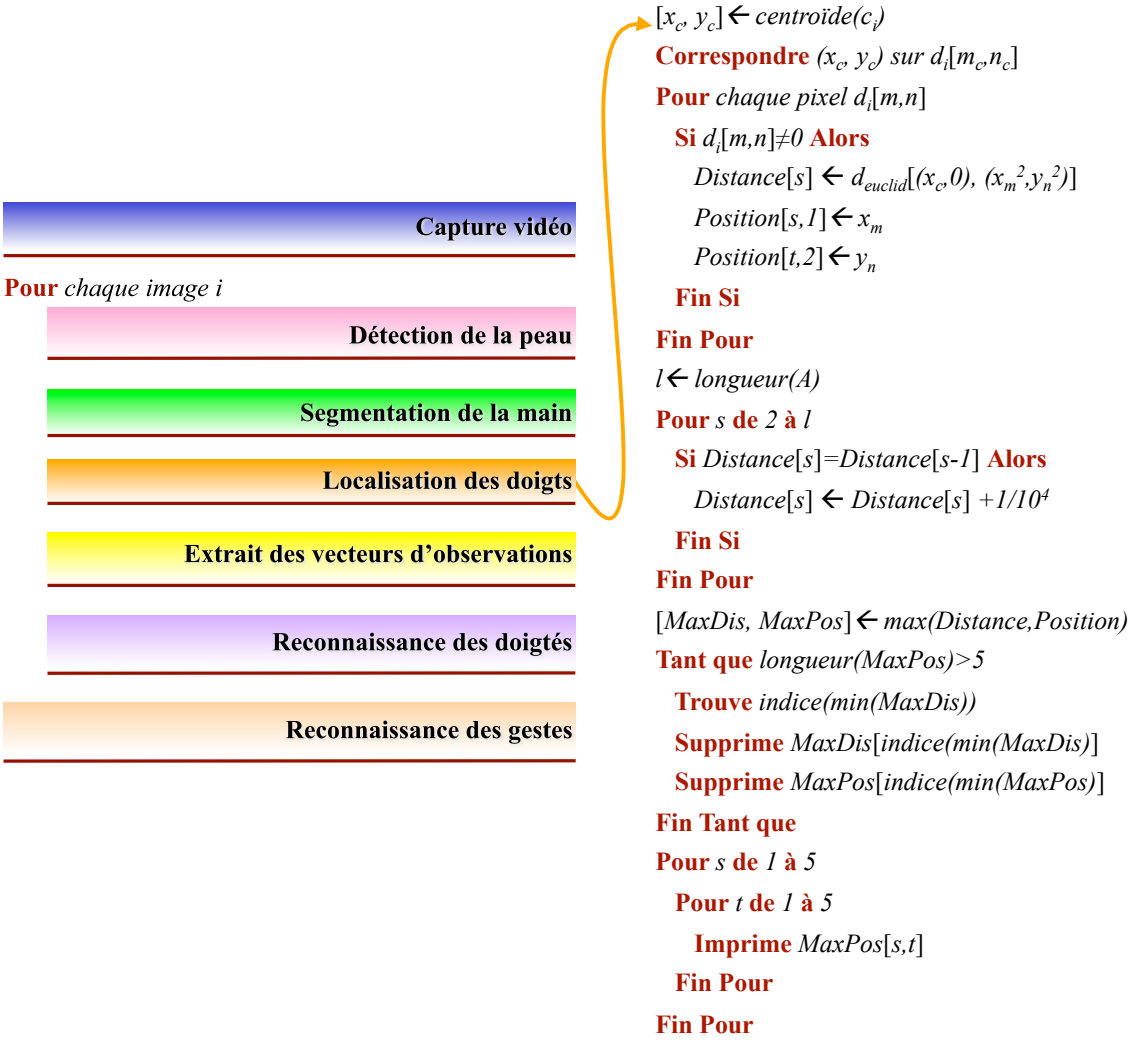

Figure 4.3.1 Algorithme de localisation des doigts dans l'image

4.5 Reconnaissance des doigtés

L'extraction (*reconnaissance*) des doigtés, autrement dit reconnaissance statique, se met en œuvre en déterminant le seuil d'appui effectué sur une touche pour chaque doigt. Même dans le cas d'un «doigt caché», le doigté sera extrait sans délai dans les images suivantes.

4.6 Reconnaissance des gestes

La combinaison des doigtés forme un geste dit «pianistique». Pour cela un dictionnaire des gestes ainsi qu'un alphabet des doigtés ont été créés. Les gestes, se projetant en mouvements musicaux, sont analysés à la fois harmoniquement et mélodiquement afin d'extraire leurs états structurels.

Les valeurs continues des vecteurs d'observation, extraits par les séquences d'images, sont modélisées à l'aide des Modèles Gaussiennes Mixtes (MGM), tandis que chaque geste est modélisé par les Modèles de Markov Cachés (MMC) [Bakis 1976; Baum 1972], offrant ainsi une certaine flexibilité à l'entraînement des modèles et permettant l'importation de vidéos de longueurs différente ou de données manquantes [Alani 1994a, 1994b].

Plus précisément, les MMC continus ont été choisis du fait (a) de la précision fournie dans la classification ; (b) qu'ils ne nécessitent pas de quantification des données ; (c) du petit nombre de données d'entraînement pour les modèles [Rabiner 1989]. Le modèle du geste est évalué en estimant le maximum de vraisemblance (*similarité entre le geste effectué et les gestes modélisés*).

4.7 PianOrasis : système et interface

PianOrasis met en oeuvre la méthodologie développée pour la reconnaissance, statique et dynamique, des gestes musicaux des doigts. Le système, ainsi que son interface, ont été entièrement développés sous Matlab. Plusieurs boîtes à outils ont été utilisées, telles que «Image Acquisition» pour la capture des vidéos, «DIPimage» pour le traitement statistique de l'image et le «Kevin Murphy» pour la modélisation stochastique à l'aide des MMC et des MGM.

PianOrasis assure différentes fonctions concernant l'importation et le suivi de traitement de la vidéo, le filtrage, l'entraînement et la reconnaissance. Plus exactement, les fonctions de PianOrasis sont présentées ci-dessous, ainsi que dans les zones (z) de la figure 4.7.1 :

- Importation d'une image/vidéo (z:1) : (a) «Open image» : ouverture d'une image ; (b) «Frame sequence» : ouverture d'une séquence d'images enregistrées sous la forme de préfixe, p. ex. : mozart0001.bmp, mozart0002.bmp etc.
- Filtrage (z:9) : «Filtering parameters» : modification des paramètres des filtres Open, Min, Gauss ainsi que ceux de la mé-

Figure 4.7.1. Interface du PianOrasis

thode de détection des contours de Canny.

- Suivi de traitement de l'image : (z:2) «Active frame» : Apparition de l'image initiale en RGB ; (z:3) «Hand extraction» : Apparition de l'image binaire créée par l'application du modèle de la peau ; (z:4) «Image filtering» : Apparition de l'image binaire de la silhouette de la main à partir de l'application des techniques de morphologie mathématique ; (z:5) «Contour» : apparition de l'image binaire des pixels du contour de la main ; (z:6) «Possible edges» : apparition des arêtes possibles correspondant aux bouts des doigts ; (z:7) «Finger detection» : Apparition des positions des bouts des doigts dans l'image avec option de prévision ((z:11) «Find missing fingers»).
- Entraînement (z:13) : «Read data from sequence» : Remplissage des vecteurs d'observation ; «Clear data» : Réinitialisation des vecteurs d'observation ; «Store sequence data» : Enregistrement des vecteurs d'observation dans une matrice ; «HMM Training» : Entraînement des modèles des gestes à partir des vecteurs enregistrés ; «Num of HMMs» : Nombre de modèles des gestes ; «Num of videos» : Nombre de vidéos analysés pour le geste actuel ; «Num of frames» : Nombre d'images analysées pour la vidéo actuelle.
- Reconnaissance : (z:10) «Fingers position» : Apparition : (a) des abscisses des doigts en position «repos» avec option de réinitialisation ; (b) des abscisses des doigts en rouge et (c) des doigtés en vert ; (z:12) «Single frame» : Reconnaissance des doigtés dans une image ; (z:12) «Frame sequence» : Reconnaissance des doigtés dans une séquence d'images ; (z:13) «Run HMM» : Reconnaissance dynamique par exportation des vraisemblances des modèles pour le geste importé ; (z:13) «Logprob» : Logarithme de la vraisemblance du modèle.

5 Evaluation

L'évaluation de la méthodologie de PianOrasis a été déployée au travers de quatre expérimentations constituant chacune une étape différente d'évaluation utilisant différents extraits musicaux.

5.1 Première expérimentation

Cette expérimentation correspond à la partition de la figure 5.1.1. Elle a été réalisée dans de mauvaises conditions d'éclairage (Figure A.1). 284 images vidéos en 25 fps ont été analysées et les résultats de la localisation par doigt apparaissent dans la figure 5.1.2.

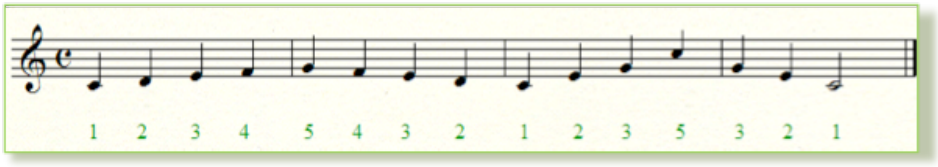

Figure 5.1.1. Premier extrait musical en DO majeur

Il est clair que pour les quatre doigts le taux de détection de leur localisation est très élevé. La physiologie du pouce entraîne un taux de détection plus faible par rapport aux autres doigts. Pourtant, ce problème peut être dépassé facilement grâce au classificateur du pouce qui prévoit sa localisation dans l'image. Dans la même vidéo, les doigtés ont été détectés à 81% dans ses images. Les doigtés de 14 des 15 notes de la partition ont été extraits correctement ; autrement dit, 93% d'entre eux ont été extraits correctement.

Image	Localisation par doigt					Localisation totale	Reconnaissance des doigtés					Note	Appui
	1er	2ème	3ème	4ème	5ème		1er	2ème	3ème	4ème	5ème		
109	0	1	1	1	1	0	0	1	1	0	0		double
110	0	1	1	1	1	0	0	1	1	0	0		double
111	0	1	1	1	1	0	0	1	0	0	0		
112	0	1	1	1	1	0	0	1	0	0	0	RE	
113	0	1	1	1	1	0	0	1	0	0	0		
114	0	1	1	1	1	0	0	1	0	0	0		
115	0	1	1	1	1	0	0	1	0	0	0		
116	0	1	1	1	1	0	0	1	0	0	0		

Figure 5.1.2. Reconnaissance statique des gestes par PianOrasis pour les images 109 à 116

Localisation : « 0 » signifie que le doigt n'est pas localisé sur l'image ; « 1 » signifie que le doigt est localisé sur l'image.
Reconnaissance : « 0 » signifie qu'il n'y a pas de détection de doigté ; « 1 » signifie que le doigté est détecté.

Exemple : Dans l'image 111 a) le 1er doigt n'est pas détecté, b) les doigts 2 à 5 sont localisés, c) le 2ème doigt appuie sur la note RE

La 15ème note (FA) n'a pas été extraite correctement, soit à cause d'un mauvais réglage du seuil d'appui pour l'annulaire, soit parce que ce doigt était caché derrière une touche noire et, par conséquent, n'a pas été localisé correctement. De toute manière, le problème peut être outrepassé en modifiant le seuil d'appui pour ce doigt.

5.2 Seconde expérimentation

La deuxième étape d'évaluation se réfère à un extrait de la «***Sonate pour piano no 16 en do majeur***» de Wolfgang Amadeus Mozart (1975-1791) (Figure 5.2.1). Une vidéo en 19 fps a été prise afin d'évaluer le système PianOrasis en utilisant moins d'échantillons par seconde, à un tempo bien plus rapide et dans de meilleures conditions d'éclairage par rapport à la première expérimentation.

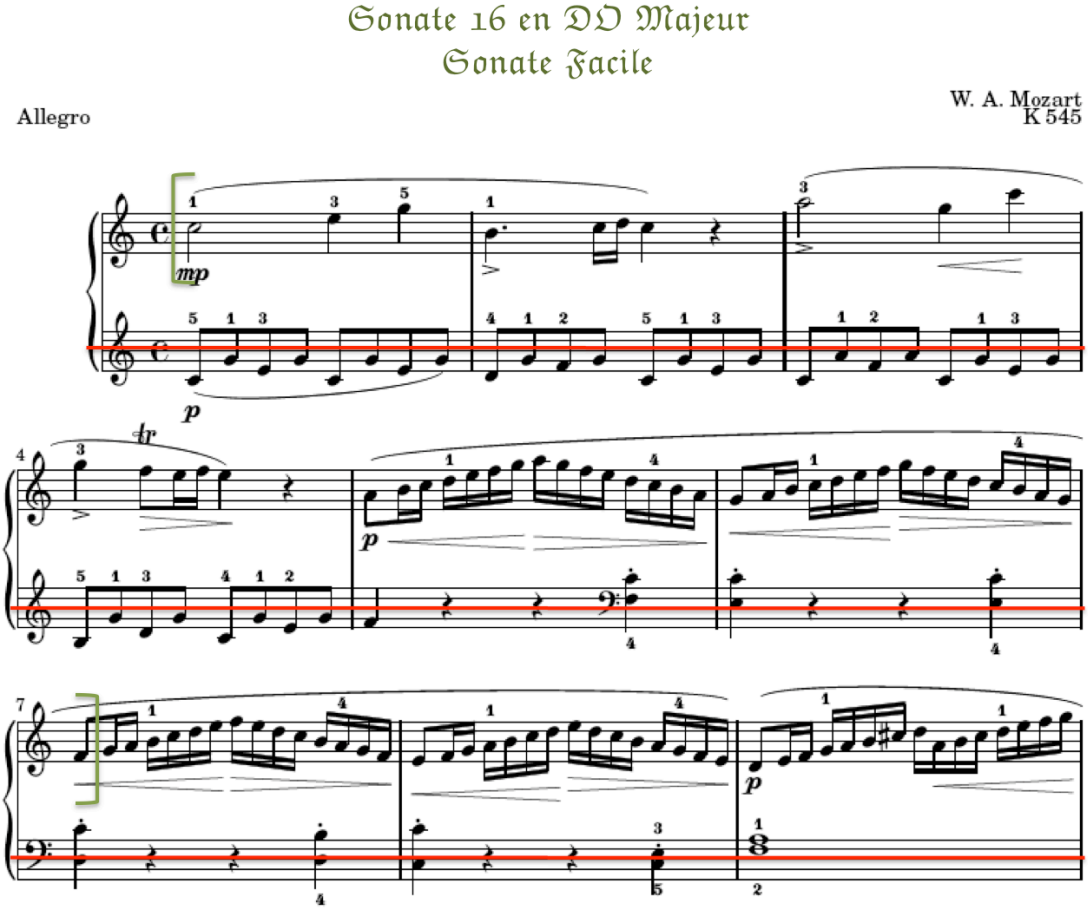

Figure 5.2.1. Deuxième extrait musical : sonate No.16 de Wolfgang Amadeus Mozart

PianOrasis présente un dysfonctionnement lorsqu'un mouvement de la main est provoqué par un saut mélodique, accompagné d'un silence dans la partition. Dans ce cas, de faux doigtés sont extraits, du fait de l'augmentation de la distance entre les bouts des doigts et le centroïd, dépassant ainsi le seuil prédéfini.

Un autre cas intéressant à citer concerne la contribution des classificateurs pour la prévision de la localisation des doigts. Dans les images 287 à 289 le pouce n'est pas détecté (*rectangle en vert, figure 5.2.2*), probablement parce qu'il est caché derrière un autre doigt. Pourtant, grâce à l'opération de prévision de la localisation des doigts pour la triade précédente des images 284 à 286, le doigté est extrait correctement (*rectangle en jaune*). Il est à noter qu'il est extrêmement rare que l'information d'un doigt caché derrière un autre, appuyant en même temps sur une touche, ne soit enregistrée dans aucune image.

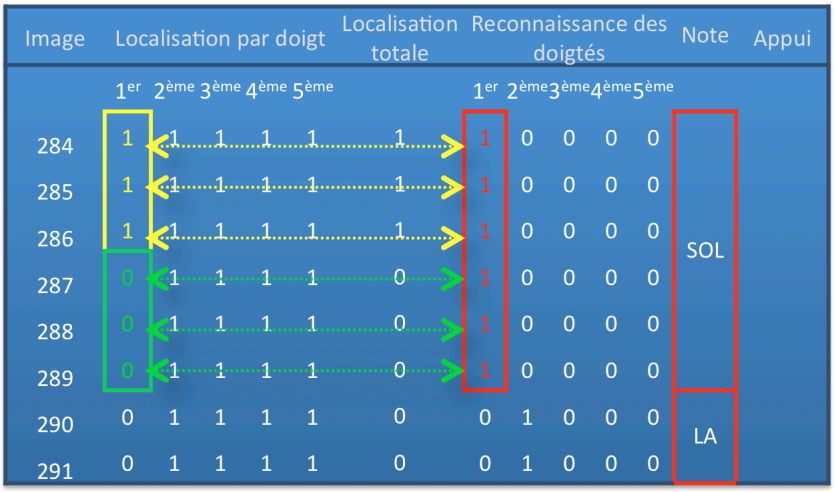

Figure 5.2.2. Reconnaissance statique des gestes par PianOrasis pour les images 284 à 291

L'étude de la figure 5.2.2 ainsi que sa comparaison avec la figure 5.2.1 peuvent fournir des conclusions intéressantes pour la qualité de la détection des doigts sans classificateurs. Cette étude révèle l'influence importante de l'éclairage dans la reconnaissance des gestes musicaux à travers une séquence d'images. La normalisation de la RI rend PianOrasis moins dépendant des variations d'éclairage mais cela ne peut en aucun cas être interprété comme une suppression des lois physiques : *un bon éclairage est meilleur qu'un mauvais éclairage*. Par exemple, le taux du pourcentage de la détection du pouce a augmenté de 12% dans la deuxième expérimentation par rapport à la première (Figure 5.2.3). Cela est dû au fait que l'éclairage dans le deuxième cas est bien meilleur que celui dans le premier, af-

fectant ainsi la détection du pouce, mais pas vraiment celle du reste des doigts. Néanmoins, la détection des doigts, excepté celle du pouce, n'a pas de variance importante grâce à leur physiologie similaire et elle reste toujours supérieure à 82% puisque le risque qu'un de ces doigts soit caché par un autre est très faible.

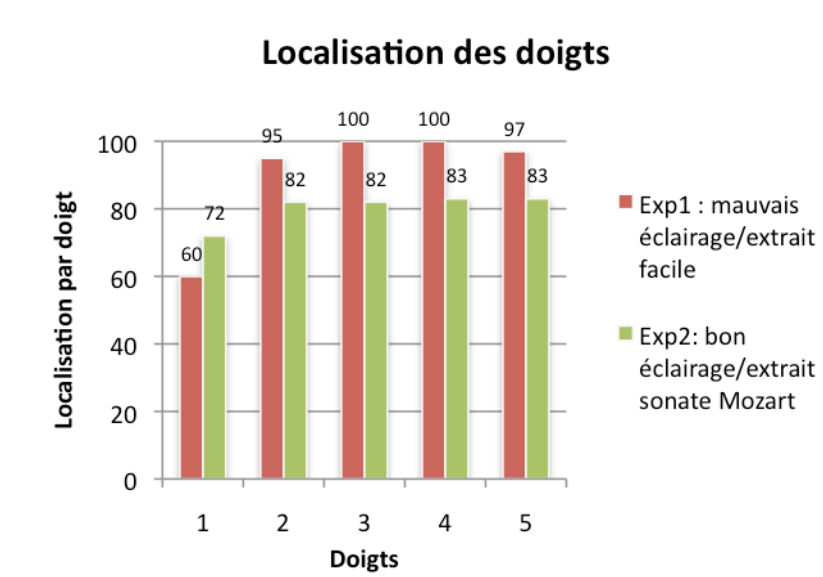

Figure 5.2.3. Comparaison de la localisation des doigts entre les deux premières expérimentations

5.3 Troisième expérimentation

La troisième étape d'évaluation consiste en une reconnaissance des gestes musicaux sans instrument musical (Figure A.2). Il a été demandé au pianiste d'interpréter l'arpège ascendant de la gamme de DO en bougeant ses doigts dans l'espace. Dans les 130 images qui ont été évaluées, la localisation des doigts, par détection et classificateurs, a été réussie à 100%, tandis que les doigtés ont été extraits correctement à 97%.

Si l'absence de touches dans l'image a provoqué des pourcentages de reconnaissance si élevés, cela est dû au fait que la silhouette de la main est entièrement détectée dans l'image. Les résultats obtenus lors de la dernière expérimentation indiquent que PianOrasis peut être utilisé, avec peu de modifications, comme un nouvel instrument numérique pour la composition de musique contemporaine en établissant une correspondance en temps réel entre les gestes reconnus et des sons modélisés. Pourtant, PianOrasis a été développé afin de reconnaître des gestes effectués sur une interface (*clavier etc*). Quand le musicien effectue le même geste dans l'espace, il se

sent obligé de respecter les seuils définis par le jeu pianistique. Ce problème peut être résolu en définissant un point de référence dans l'image et en calculant ainsi toutes les distances par rapport à ce point.

5.4 Quatrième expérimentation

Au cours de la quatrième expérimentation, le système PianOrasis a été entraîné à reconnaître les gestes pianistiques présentés autour de la gamme de DO : (a) la gamme ascendante (GA) ; (b) la gamme descendante (GD) ; (c) l'arpège ascendant (AA) ; (d) l'arpège descendant (AD) ; (e) les tierces ascendantes (TA) et (f) les tierces descendantes (TD).

PianOrasis a été entraîné sur les 14 différents vecteurs d'observation, à l'aide de 120 vidéos en 19 fps (*20 pour chaque geste modélisé*). Le nombre des états du modèle du geste a été défini en fonction du nombre total des appuis (doigtés) effectués mélodiquement sur les touches. Par exemple, le modèle de l'arpège ascendant a été modélisé suivant 4 états. L'évaluation de PianOrasis dans ce scénario consiste à reconnaître des « gestes isolés » dans une séquence d'images. Il a été demandé au système de reconnaître chaque geste dans 10 vidéos différentes. Les résultats de cette étape d'évaluation sont présentés dans la figure 5.4.1.

Les taux de reconnaissance pour les gestes TA et TD sont les moins élevés parmi tous les gestes. Ces deux gestes sont assez complexes ayant un niveau stochastique très élevé. Par conséquent, plus le geste est stochastique, plus le nombre nécessaire de vidéos d'entraînement augmente et plus la reconnaissance devient difficile. Dans huit vidéos différentes, les modèles du GA et du GD ont eu la vraisemblance maximale pour les gestes TA et TD et vice versa. Il existe une forte similarité entre les paires de gestes TA/GA et TD/GD, car tous les quatre gestes ont été modélisés suivant le même nombre d'états et les positions de départ de la main/centroïd, ainsi que celles d'arrivée, sont très proches l'une de l'autre pour les deux paires des gestes.

Dans le cas des gestes AA et AD, les taux de reconnaissance sont très élevés étant donné qu'il s'agit de gestes « simples » ayant un niveau stochastique bas. Un autre élément qui prouve la simplicité de ces deux gestes est le petit nombre d'appuis effectués mélodiquement sur les touches, provoquant un nombre d'états assez réduit par rapport aux autres gestes modélisés, ainsi que des vidéos d'entraînement très courtes.

gestes	*GA*	*GD*	*AA*	*AD*	*TA*	*TD*	*non classifiés*
GA	7	0	0	0	2	0	1
GD	0	8	0	0	0	2	0
AA	0	0	10	0	0	0	0
AD	0	0	0	10	0	0	0
TA	2	0	0	0	6	0	2
TD	0	2	0	0	0	5	3
reconnaissance	**70%**	**80%**	**100%**	**100%**	**60%**	**50%**	**10%**

Figure 5.4.1. Reconnaissance dynamique par PianOrasis

Croisement rang/colonne : nombre de fois où le modèle du geste (colonne) a eu la vraisemblance maximale pour le geste importé (rang) dans PianOrasis ;

Gestes non classifiés : les vraisemblances ont été inférieures à un seuil pour tous les modèles.

6 Contribution

La thèse présentée couvre une large gamme de champs d'application, touchant plusieurs domaines scientifiques aussi divers que celui de la musique, de l'informatique ou encore de la mise en valeur du patrimoine culturel.

6.1 Musique

La contribution de la méthodologie présentée dans la musique repose sur deux axes principaux : (a) la pédagogie musicale et (b) la composition de musique contemporaine.

PianOrasis peut contribuer à la pédagogie musicale en tant que support informatisé pour l'apprentissage du piano. La technique des doigts est pour un pianiste l'alphabet de son interprétation musicale. L'étude et l'observation des doigts du pianiste, pendant son interprétation, ainsi que de la manière dont il interagit avec le piano, constituent une méthode excellente pour la compréhension approfondie de l'œuvre musical et des sentiments du pianiste. Dans ce cadre, PianOrasis peut être utilisé en tant que

système d'optimisation de la technique des doigts et du choix des doigtés au piano.

Les résultats obtenus, lors de l'expérimentation de la reconnaissance des gestes musicaux sans instrument musical, indiquent que PianOrasis peut être utilisé, avec peu de modifications, comme un nouvel instrument numérique pour la composition de musique contemporaine. Cela peut être réalisé en mettant en relation deux domaines distincts, que sont le geste et le son. Dans ce cadre, PianOrasis peut servir à une large gamme de scénarii de composition de musique contemporaine, en tant qu'une interface tangible pour la reconnaissance des gestes musicaux effectués sur un objet ou une surface dans un environnement réel.

6.2 Mise en valeur du patrimoine culturel immatériel

Le patrimoine culturel matériel est le fruit des savoir-faire de haute technicité, constitués et transmis au fil des siècles grâce à l'intelligence du geste associée à la créativité de l'esprit humain. Jusqu'à ce jour, les gestes des artisans ne se sont jamais laissés «mettre en boîte», enregistrer, classifier, codifier de manière à pouvoir être transmis, même après leur extinction, par quelque moyen que ce soit. Par conséquence, la méthodologie proposée dans cette thèse, ainsi que sa mise en œuvre pour la reconnaissance et la modélisation des interactions gestuelles entre les artisans et leur matière, consisterait une innovation dans le domaine de la sauvegarde des savoir-faire rares.

6.3 Interaction Homme-Machine

La similarité entre les gestes effectués sur un clavier de piano et ceux effectués sur un clavier d'ordinateur est forte sans aucun doute. Une telle méthodologie de VpO peut favoriser le remplacement des périphériques restrictifs d'un ordinateur, tels que le clavier ou la souris, par une caméra embarquée à l'écran de l'ordinateur et un algorithme de reconnaissance des mouvements articulés des doigts, en palmier semi-étendu pour chaque doigt. A ce point, il serait utile de citer la contribution de la méthode à la disparition des maladies liées aux claviers, tel que le syndrome du canal carpien.

6.4 Sciences de la vie et de la santé

La méthodologie proposée peut être appliquée dans la réadaptation fonctionnelle de la main, voire l'interaction du patient avec son environnement physique et réel. Par exemple, la modélisation des gestes de réadaptation ainsi que l'entraînement du système sur ces gestes, contribuerait au suivi des doigts du patient exerçant des mouvements précis d'une difficulté individualisée.

Le système, en tant qu'une Interface Homme-Machine, peut apporter une contribution très importante à l'étude et au décodage du comportement humain au travers des gestes, ainsi qu'à l'assistance à la vie en autonome, surtout pour les personnes handicapées. Finalement, une telle interface contribue également au développement d'outils technologiques pour l'acquisition multimodale de données des patients dans la recherche médicale.

7 Conclusions

7.1 Originalité

L'objectif de la thèse présentée ici était le développement d'une méthodologie de vision par ordinateur pour la reconnaissance des gestes musicaux des doigts à l'aide de l'analyse et de la modélisation stochastique du geste.

Dans la première section du synopsis étendu, les critères que la méthodologie devait suivre, ont été définis. La méthodologie présentée ici satisfait les critères de la première section puisqu'elle est :

1. **Capable** : elle est capable de calculer tous les paramètres définissant les gestes des doigts, indépendamment du style d'interprétation musicale, suivant un entraînement ;
2. **Vision orientée** : elle est orientée vers l'image du musicien, sans analyse préliminaire de la partition ;
3. **Non intervenante** : le musicien est complètement libre à interpréter et s'exprimer musicalement avec des gestes des doigts, soit sur une interface soit dans l'espace, sans aucun équipement spécifique ;
4. **Accessible** : elle est à faible coût, permettant la reconnaissance des gestes des doigts dans plusieurs champs d'applications et à grande échelle.

7.2 Restrictions et perspectives

PianOrasis a certaines restrictions techniques. Ces restrictions se présentent ci-dessous :

1. **Caméra :** a) La reconnaissance ne s'effectue que pour les touches blanches ; b) Le plan de la capture se limite à deux octaves du piano ; c) La modification de la résolution de l'image nécessite l'adaptation des vidéos d'entraînement ;
2. **Capture unimodale de données :** Les données dans la séquence d'images concernent uniquement la main droite du pianiste ;
3. **Reconnaissance partielle limitée :** a) PianOrasis ne reconnaît pas les gestes qui s'effectuent sur les touches noires ; b) PianOrasis ne reconnaît pas en temps réel.

Ces limites ouvrent autant de perspectives de recherche :

1. **Caméra :** a) captage de vision large ; b) support spécifié pour la caméra ;
2. **Acquisition multimodale de données :** importation des sons dans PianOrasis et correspondance avec les gestes modélisés ;
3. **Reconnaissance pleine en temps réel :** reconnaissance en temps réel des gestes effectués sur n'importe quelle surface ou objet ;
4. **Expression musicale :** production du son à partir des gestes reconnus dans l'espace,

7.2 Motivations et résultat

Motivés par l'absence de liens opérationnels entre les domaines de la vision par ordinateur et de l'interaction musicale, nous avons étudié dans cette thèse la conception et la mise en oeuvre d'une méthodologie et d'un système de vision par ordinateur au service de l'expression musicale libre, en obtenant ainsi l'objectif de la reconnaissance des gestes musicaux des doigts sans interférence entre le musicien et l'ordinateur.

Bibliographie

[Alani 1994a]

Alani, T., Guelif, H. (1994). *Modèles de Markov Cachés - Aspects pratiques*. INRIA, France.

[Alani 1994b]

Alani, T. (1994). *Modèles de Markov Cachés - Théorie et techniques de base*. ESIEE, France.

[Albrecht 2003]

Albrecht, I., Haber, J. & Seidel, H. P. (2003). Construction and animation of anatomically based human hand models. In *Proceedings of the 2003 ACM SIGGRAPH/Eurographics symposium on Computer animation,* San Diego, California, USA.

[Bakis 1976]

Bakis, R. (1976). Continuous speech recognition via centisecond acoustic states. *The Journal of the Acoustical Society of America, New York*, *59*(1), 97.

[Baum 1972]

Baum, L. (1972). An inequality and associated maximization technique in statistical estimation for probabilistic functions of Markov processes. In *Proceedings of the Third Symposium on Inequalities,* New York, USA.

[Bérard 2000]

Bérard, F. (2000). « *Vision par Ordinateur pour l'interaction homme-machine fortement couplée* », Thèse de doctorat, Université de Joseph Fourier, Grenoble, France.

[Bevilacqua 2007]

Bevilacqua, F., Guédy, F., Schnell, N., Fléty, E., Leroy, N. (2007). "Wireless sensor interface and gesture-follower for music pedagogy", In *Proc. of the International Conference of New Interfaces for Musical Expression (NIME 07)*, p 124-129.

[Bevilacqua 2010]

Bevilacqua, F., Zamborlin, B., Sypniewski, A., Schnell, N., Guédy, F., Rasamimanana, N. (2010). Continuous realtime gesture following and recognition, *LNAI 5934*, pp. 73–84.

[Boukir 2004]

Boukir, S. & Chenevière, F. (2004). « Conception d'un système de reconnaissance de gestes dansés». *Traitement du signal, 21*(3), 195-203.

[Burns 2006]

Burns, A. M. & Wanderley M. (2006). Visual Methods for the Retrieval of Guitarist Fingering. In *Proceedings of the International Conference on New Interfaces for Musical Expression*. Paris, France.

[Cadoz 2000]

Cadoz, C., & Wanderley, M. M. (2000). Gesture - music [electronic]. In M. M. Wanderley & M. Battier (Eds.), *Trends in gestural control of music*, 29–65, IRCAM.

[Coduys 04]

Coduys, T., Henry, C. and Cont, A. (2004). "TOASTER and KROONDE: High-Resolution and High-Speed Real-time Sensor Interfaces", In *Proc. of the International Conference on New Interfaces for Musical Expression (NIME-04)*, Hamamatsu, Japan.

[Decroux 1994]

Decroux, E. (1994). *Paroles sur le mime*. Librairie Théâtrale : Paris.

[Demoucron 08]

M. Demoucron, A. Askenfelt, and R. Caussé. (1994). Observations on bow changes in violin performance. In *Proceedings of Acoustics, Journal of the Acoustical Society of America*, volume 123, page 3123.

[Grunberg 2008]

Grunberg, D. (2008). *Gesture Recognition for Conducting Computer Music*. Retrieved January 10, 2009, from: http://schubert.ece.drexel.edu/research/gestureRecognition

[Palmer 2000]

Palmer, C. & Pfordresher, P. Q. (2000). From my hand to your ear: the faces of meter in performance and perception. In C. Woods, G. Luck, R. Brochard, F. Seddon & J. A. Sloboda (Eds.) In *Proceedings of the 6th International Conference on Music Perception and Cognition*. Keele, UK: Keele University.

[Papamarkos 2000]

Papamarkos, N., Strouthopoulos, C., & Andreadis, I., (2000). "Multithresholding of color and gray level images through a neural network technique", *Image and Vision Computing*, vol. 18, 213-222.

[Rabiner 1989]

Rabiner, L. R. (1989). «A tutorial on hidden Markov models and selected applications in speech recognition». *Proceedings of the IEEE, 77(2),* 257-285.

[Radicioni 2004]

Radicioni, D., Anselma, L. & Lombardo, V. (2004). An Algorithm to compute fingering for string instruments. In *Proceedings of the 2nd national congress of the associazione italiana di scienze cognitiva*. Ivrea, Italy.

[Rasamimanana 2008]

Rasamimanana, N., Bernardin, D., Wanderley, M. & Bevilacqua., F. (2008). String bowing gestures at varying bow stroke frequencies: A case study. In Advances in Gesture-Based Human-Computer Interaction and Simulation, *volume 5085 of Lecture Notes in Computer Science, pages 216–226. Springer Verlag*, 2008.

[Rasamimanana 2009]

Rasamimanana, N. & Bevilacqua., F. (2009). Effort-based analysis of bowing movements: evidence of anticipation effects. *The Journal of New Music Research,* 37(4):339 – 351, 2009.

[Traube 2004]

Traube, C. (2004). *An interdisciplinary study of the timbre of the classical guitar*. Unpublished doctoral dissertation, McGill University.

[Verner 1995]

Verner, J. A. (1995). « MIDI guitar synthesis yesterday, today and tomorrow, an overview of the whole fingerpicking thing », *Recording Magazine*, *8*(9), 52-57.

[Vicon Peak 2005]

Vicon Peak. (2005). *Vicon Motion Capture System*, Lake Forest, CA.

Annexe A : Copies de l'interface du PianOrasis

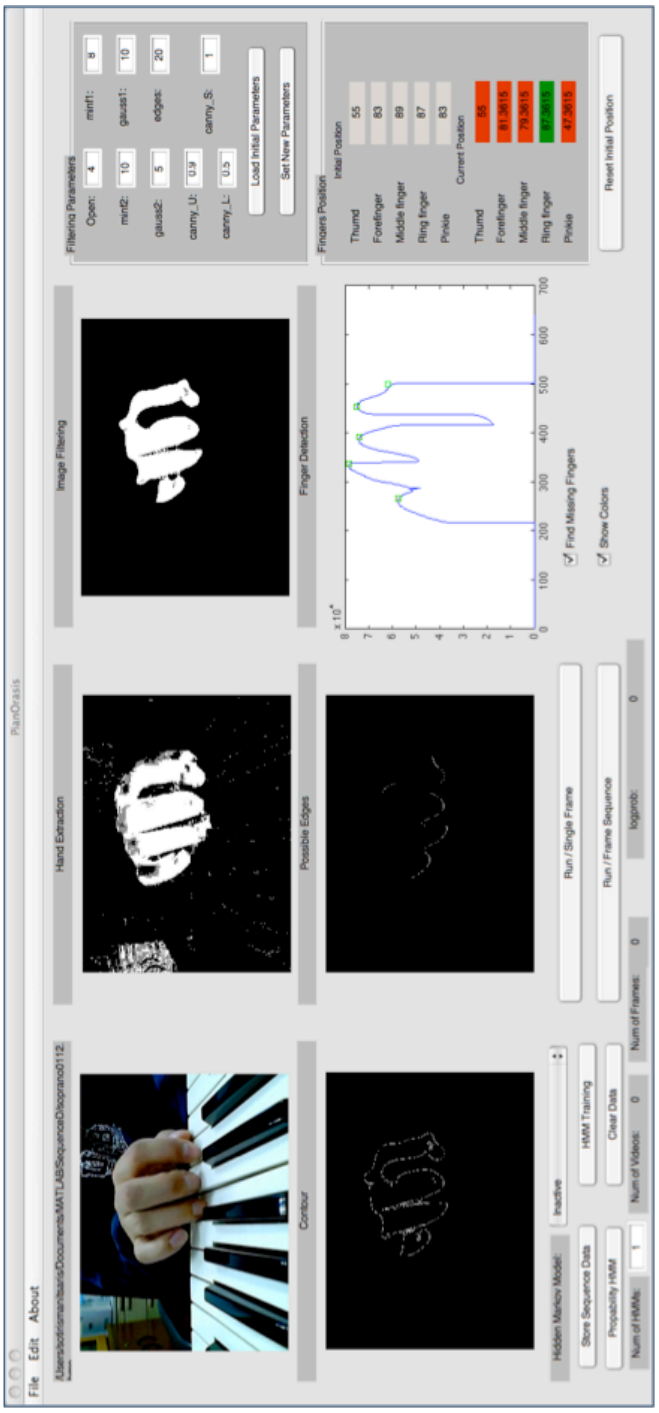

Figure A.1. PianOrasis et traitement de l'image 112 (*première expérimentation*)

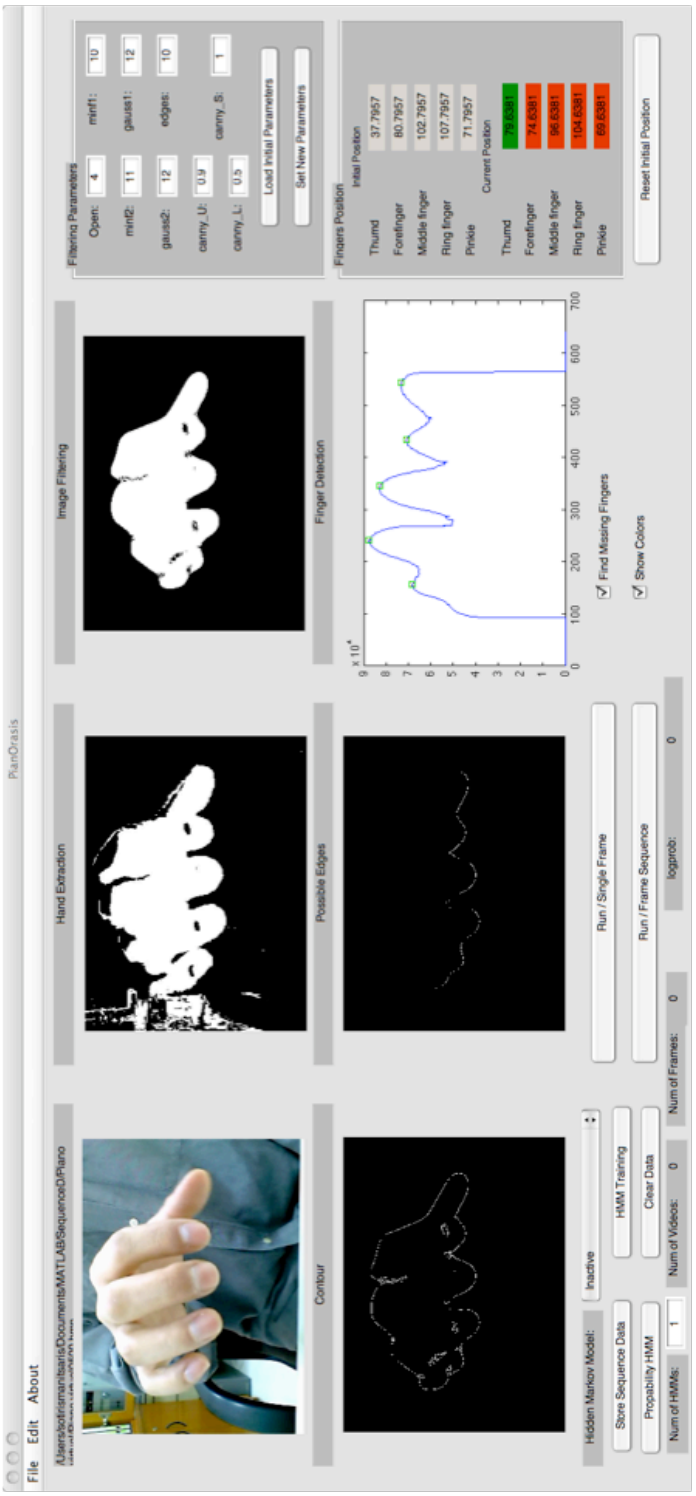

Figure A.2. PianOrasis et reconnaissance des gestes musicaux des doigts dans l'espace