Interactions geste-musique par vision artificielle
Gesture-music interactions by artificial vision

F. Bardet T. Chateau F. Jurie M. Naranjo

Lasmea, UMR6602 du CNRS/Université Blaise Pascal, 63172 Aubiere Cedex, France
thierry.chateau@lasmea.univ-bpclermont.fr

Résumé

Nous développons un système d'interaction homme-machine, permettant de produire de la musique à partir de la vision par ordinateur, du geste humain. La solution présentée suit les gestes des deux mains d'une personne, à partir d'un flux vidéo couleur temps réel, pour déclencher des événements sonores et pour contrôler continument des paramètres du son. Des gestes codifiés font évoluer le système dans une séquence prédéterminée, modélisée sous forme de machine d'état. Les applications visées par ces travaux sont, entre autres, le spectacle vivant ou l'enseignement.

Mots Clef

Interaction Homme-Machine, Musique, Suivi d'objets.

Abstract

We are developping a human-machine vision-based interaction system, allowing the user to generate music from his gestures. The prototype we present here, tracks and analyses both hand gestures of a human, from a single camera in real time, in order to trigger sound events and to continuously change sound parameters. Some known gestures allow the system to evolve along a sequence. The application we are developping shows interesting possibilities in live performances and in education.

Keywords

Human-Machine interaction, Music, Visual tracking

Introduction

L'objectif de cette recherche est l'étude et la réalisation d'une plate-forme économique et portable, assurant le contrôle gestuel en temps réel, de l'environnement sonore d'un spectacle vivant. Le produit final permettra par exemple à un mime, un danseur, ou un marionnettiste, de jouer la musique accompagnant sa prestation. Le danseur générera et contrôlera ainsi par ses gestes la structure rythmique sonore, au lieu de la suivre comme c'est généralement le cas. Dans une autre application, la réalisation finale permettra de monter des installations d'interaction gestuelle, dans lesquelles le public jouera lui-même avec le son. Il devient alors un acteur en interaction avec la machine.

Nous décrivons ici le résultat de la première étape de ce travail. Ce prototype prend en compte les positions 2D des deux mains d'une personne, observée par un système d'acquisition vidéo monoculaire, pour réaliser deux actions : le traitement et restitution temps réel d'une phrase musicale vocale enregistrée ; un pseudo-instrument virtuel qui permet de poser des notes de contrebasse en contrepoint à ce chant. Des traitements simples permettent de reconnaître des gestes tels que prendre, poser... Le système fonctionne sur un ordinateur de type PC exploitant l'image couleur de la scène, et délivrant le signal sonore sur sa sortie audio.

Après une présentation des gestes et de leurs interactions dans le spectacle vivant, ainsi que des objectifs de cette étude, un état de l'art des systèmes d'interaction geste-son est proposé dans une première partie. La deuxième partie présente les méthodes de traitement d'images, d'analyse de gestes et de modélisation de l'interaction geste-son. Une troisième partie traite des aspects techniques liés à l'acquisition des images et à la génération du son par ordinateur. La dernière partie conclue et propose des extentions possibles à ce travail.

1 Gestes et interactions dans le spectacle vivant

On appelle geste tout mouvement ou posture du corps ou d'une partie du corps. Un geste sera dit expert, s'il est le produit d'un apprentissage et d'une adaptation aux contraintes physiques d'un outil (instrument de musique par exemple). Le geste expert est un geste qui s'est adaptaté, au cours de l'apprentissage, aux contraintes physiques de l'outil. Il a acquis du savoir-faire, mais en contrepartie, a perdu de sa spontanéité.

On appelle interaction toute action mutuelle réciproque de deux agents (phénomènes ou personnes). Le spectacle musical repose sur les interactions entre le public, les musiciens, et leurs outils (les instruments de musique).

– Il y a toujours interactions inter-scénique (entre les mu-

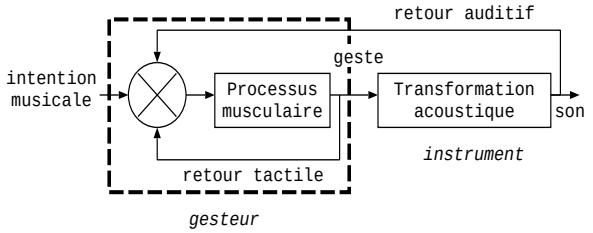

FIG. 1 – retours du geste instrumental.

siciens, entre le musicien et son instrument) et extra-scéniques (entre les musiciens et le public) mais le retour du public vers la scène est peu développé.
– L'instrumentiste exerce un geste expert sur son instrument. L'instrument reçoit, traite, et transmet à l'air cette excitation mécanique, pour la rendre audible. Cette chaîne d'action est contrôlée par deux boucles de retour : un retour tactile, et un retour auditif. Ce schéma est présenté sur la figure 1. On appelle "geste instrumental" un geste exécuté en contact avec une matière, par opposition au "geste à main nue". Des études comme celles de C. Cadoz [2] ont montré que l'instrumentiste ne peut se passer du retour tactile.
– L'informatique permet de nouveaux types d'interactions. Quelle peut être la nature de ces interactions ?
– En musique, depuis le Theremin (1922) - instrument sensible à la variation du champ électrique produit par la personne humaine - un grand nombre de recherches concerne les instruments augmentés, et les instruments virtuels. Les instruments augmentés sont des instruments de musique conventionnels, augmentés de possibilités de jeu (donc des possibilités d'expression) plus étendues. En contre partie l'expertise gestuelle requise est augmentée d'autant. La conception d'instruments virtuels (les nouvelles lutheries) constitue aussi une très forte attraction. Mais ils doivent synthétiser un retour tactile, sans quoi le geste n'est pas instrumental. Le Clavier Rétroactif Modulaire de l'ACROE [2] réalise ce retour au prix d'une mécanique complexe.

1.1 Orientations de nos travaux : interactions avec un geste à main nue

Pour les raisons exposées plus haut, notre choix s'est porté sur l'interaction gestuelle à base de gestes à main nue (non instrumentaux), et de niveaux d'expertise variables.

Nous cherchons à étendre l'interaction à d'autres champs que l'interaction gestuelle instrumentale :
– Un premier champ est celui où opère un gesteur expert, mais dont la fonction principale du geste n'est pas la production sonore (théâtre, danse, mime, marionnette...). Ce gesteur expert a travaillé son geste en vue d'un résultat à dominante visuelle. Si nous augmentons son espace d'expression par la possibilité d'un contrôle musical, il mettra au point de même un contrôle expert de la musique avec ses gestes. Ce travail préalable lui permettra de prévoir et contrôler parfaitement par ses gestes, le son qui l'accompagne pendant le spectacle. Dans cette logique, le gesteur est au premier plan, et l'accompagnement musical est asservi à ses gestes en temps réel.
– Le deuxième champ fait intervenir un gesteur non expert dans des installations sonores interactives. Ses gestes sont plus spontanés, non travaillés, et il doit découvrir les relations entre gestes et sons. L'intérêt de cette situation est la découverte par le gesteur, des relations geste-musique, choisies par le concepteur. Mais sa gestuelle non experte ne lui offre qu'une capacité d'expression réduite. Le résultat musical serait donc pauvre s'il en était le seul agent. C'est là que l'interaction sera fructueuse, si elle fait coopérer l'homme et la machine dans la production d'un résultat musical. La machine doit donc être douée d'un minimum de savoir-faire musical, pour par exemple, répondre à une improvisation du gesteur, par une autre improvisation. L'enjeu est de faire participer un non instrumentiste à une expérience d'écoute et d'expression musicale.
– Ce champ d'interaction avec le gesteur non expert permet des applications pédagogiques, permettant au non instrumentiste ou au futur instrumentiste, de développer sa créativité musicale par interface gestuelle, sans se poser de problème de technique instrumentale. L'idée est de mettre à sa disposition une possibilité d'expression musicale, avant même qu'il ne soit instrumentiste.
– Autre application : un système thérapeutique aidant les personnes handicapées à éduquer ou à rééduquer leurs facultés motrices ou sensorielles à l'aide du son. Exemple : faire entendre à un aveugle un son venant de la direction qu'il pointe du doigt.

1.2 Etat de l'art

On peut classifier les systèmes d'interaction geste-son en fonction de la nature des capteurs utilisés.

Acquisition du geste par contact Dans la plupart des applications actuelles contrôlant le son, le geste est mesuré par des capteurs posés à même le corps : Axel Mulder revêt un danseur, d'une combinaison instrumentée [13], pour interagir avec le son produit par un violoniste (violon MIDI), Joe Paradiso [15] du MIT Medialab, a instrumenté une chaussure pour accéder aux mouvements de celui qui la porte, Wechsler accède aux contractions musculaires et aux battements du coeur. Ces capteurs sont très performants, mais très intrusifs : il faut porter des gants ou une combinaison instrumentés, pour voir ses gestes pris en compte. Cet équipement à même le corps est souvent coûteux, peu fiable et peu transportable, et surtout, il entrave les gestes.

Acquisition du geste à distance A distance, on instrumente le gesteur d'émetteur(s) de champ magnétique, et on le repère dans l'espace, par des capteurs fixes. Le bâton digital de Max Mathews [12] est une percussion augmentée : les baguettes de percussion sont équipées d'un émetteur radio, et son . On déclanche ainsi des commandes musicales, en plus du jeu de percussion classique. Les cap-

teurs de champ magnétique sont très utilisés en réalité virtuelle ou augmentée, mais leur étendue de mesure est trop faible (qqs dizaines de cm) pour nos besoins. Il reste enfin la mesure optique avec ou sans marqueurs, aujourd'hui la plus utilisée, notamment en réalité virtuelle, analyse du geste sportif ou en animation. Le système VICON, [17]), par exemple, échantillonne à 120 Hz les informations délivrées par 8 caméras, pour reconstruire en temps réel la position 3D d'un corps sur lequel sont posés 30 marqueurs (des réflecteurs optiques). Cette installation lourde et coûteuse est peu compatible avec la scène.

Acquisition du geste par vision artificielle La recherche d'un geste naturel et spontané interdit l'usage des capteurs intrusifs. De plus, le temps réel impose que le flux de données à traiter reste modeste. Pour ces raisons nous avons choisi une solution basée sur la vision monoculaire.
Les travaux les plus représentatifs concernant l'analyse du geste non instrumental par vision artificielle en temps réel sont les suivants :

– le Laboratoire d'Informatique Musicale de l'université de Gênes a développé un environnement dédié : Eyesweb [3]. Leur approche est orientée sur la perception de l'expression à travers les mouvements de l'ensemble du corps d'un seul danseur, à l'aide de deux caméras.
– Flavia Sparacino [16] du MIT Media Lab utilise Vicon ou Pfinder [18] pour créer des "performances augmentées" croisant théâtre, musique et danse. Pfinder extrait une personne du fond en monovision, puis la modélise et la suit par tâches de couleur.
– l'université d'Irvine a développé le logiciel MCM (Motion Capture Music) qui met en relation les informations posturales fournies par VICON, avec la musique [7].

2 Solution retenue

L'approche présentée ici se décompose en trois grands blocs, représentés sur le synoptique de la figure 2. La scène est observée à l'aide d'un dispositif d'acquisition grand public (webcam). Le flot vidéo est ensuite traité dans un module d'analyse d'images dont le but est de rechercher la position des deux poignets du gesteur. La position de chaque poignet, exprimée dans le repère 2D lié à l'image est ensuite transmise à un module de reconnaissance. Ce dernier analyse les trajectoires effectuées par chaque poignet afin de reconnaître certains types de gestes. Les gestes reconnus sont alors transmis à une dernier module de traitement du son dont le but est de gérer l'interaction geste-son.

2.1 Analyse de l'image

Le suivi des mains par traitement d'images. Dans l'approche proposée ici, le gesteur est équipé de 2 gants de couleur différente, afin de faciliter la détection et le suivi des mains. La détection de la position des deux gants, à partir du flux vidéo, est assurée par un algorithme de suivi d'objets colorés.
Le suivi d'objets dans des séquences d'images est un axe de recherche très largement abordé en traitement d'images.

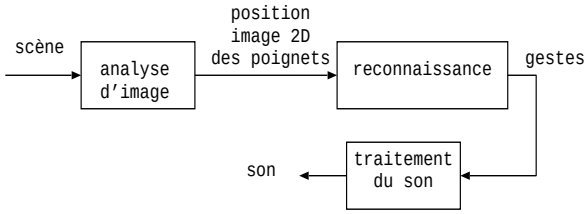

FIG. 2 – synoptique général du système.

Une classification des travaux développés autour de cette problématique est assez complexe à faire car de nombreux paramètres doivent être considérés lors du choix d'une méthode de suivi.

Certaines approches comme [6] utilisent un modèle géométrique *a priori* de l'objet. Ce dernier peut être basé sur les contours [10], la texture [9], ou sur la couleur [4, 14]. Dans certains cas, le suivi doit être particulièrement robuste aux variations d'éclairement ou aux occultations partielles [5].

Dans le cas du suivi de gants, il s'agit d'objets non rigides colorés. Il faut être particulièrement robuste aux occultations. Une maîtrise des caractéristiques du fond de la scène est un plus. Dans le cas de la méthode proposée ici, il est indispensable que les couleurs des gants soient différentes de celles du fond. La méthode proposée se divise en 2 parties : l'apprentissage de la couleur dominante de chaque gant et leur suivi.

L'apprentissage de la couleur. Le but de l'apprentissage de la couleur est de rechercher, dans une région d'intérêt $\mathtt{W}^i$ (définie de façon supervisée) de l'image $\mathtt{I}$, la couleur la plus représentative (dominante). Cette dernière est codée dans un repère HSV (Hue, Saturation, Value) car il permet de séparer l'intensité des deux autres composantes. Soit $\mathbf{y}(\mathbf{u})$ le vecteur $(h, s, v)^T,\ h, s, v \in [0, 1]$ (teinte, saturation, intensité) associé au pixel de coordonnées $\mathbf{u}$.

L'apprentissage de la couleur dominante est obtenu par un système de vote. Pour chaque pixel, on calcule une distance entre sa couleur et la couleur de chacun des autres pixels de la région d'intérêt. La somme de toutes les distances ainsi calculées forme alors un poids associé à la couleur du pixel courant. La fonction de poids $g_c^i(u)$, associée au gant i, est alors définie par :

$$\forall \mathbf{u} \in \mathtt{W}^i, g_c^i(\mathbf{u}) = \sum_{\mathbf{x} \in \mathtt{W}^i,\ \mathbf{x} \neq \mathbf{u}} \exp\left[-\lambda_c . d_c(\mathbf{y}(\mathbf{x}), \mathbf{y}(\mathbf{u}))\right] \tag{1}$$

Le paramètre λ_c permet d'ajuster le calcul du poids. La distance d_c utilisée est une distance de Mahalanobis définie par :

$$d_c(\mathbf{y}_1, \mathbf{y}_2) = [\mathbf{y}_1 - \mathbf{y}_2]^T . \mathtt{C}_c^{-1} . [\mathbf{y}_1 - \mathbf{y}_2] \tag{2}$$

avec $\mathtt{C}_c$, matrice de covariance associée aux trois compo-

santes du vecteur couleur :

$$\mathtt{C}_c = \begin{pmatrix} \sigma_h^2 & 0 & 0 \\ 0 & \sigma_s^2 & 0 \\ 0 & 0 & \sigma_v^2 \end{pmatrix} \quad (3)$$

Dans le calcul du modèle, l'intensité n'est pas prise en compte : on choisira donc $\sigma_v \rightarrow \infty$.
La couleur dominante associée à chaque gant, notée $\mathbf{y}^{i*}$ est donnée par celle du pixel qui posséde le poids maximum :

$$\mathbf{y}^{i*} = \mathbf{y}(\text{argmax}(g_c^i)) \quad (4)$$

Le suivi. Une fois le modèle de couleur associé à chaque gant ($\mathbf{y}^{i*}$) connu, il convient de rechercher, dans chaque nouvelle image, la position des deux gants. La plupart des algorithmes de suivi sont basés sur la définition d'une région d'intérêt remise à jour à chaque image. Nous avons choisi une approche différente, qui consiste à balayer toute l'image afin de rechercher le plus gros amas de points dont la couleur est proche de celle du modèle. Ainsi, une liste de points candidats $\mathtt{L}^i$:

$$\mathtt{L}^i = \{\mathbf{u} \in \mathtt{I}_s / \ |\mathbf{y}(\mathbf{u}) - \mathbf{y}^{i*}| < \boldsymbol{\sigma}\}, \boldsymbol{\sigma} = (\sigma_h, \sigma_s, \sigma_v)^T \quad (5)$$

$\mathtt{I}_s$ est constitué des points de $\mathtt{I}$ dont la valeur de l'intensité est supérieure à un seuil v_l. En effet, un point dont l'intensité est proche de zéro possède une teinte quelconque :

$$\mathtt{I}_s = \{\mathbf{u} \in \mathtt{I} / \ \mathbf{y}(\mathbf{u}) > (0, 0, v_l)^T\} \quad (6)$$

Dans les équations 5 et 6, l'opérateur d'inégalité $<$ (resp. $>$) utilisé entre 2 vecteurs est vérifié si les inégalités concernant chaque élément des 2 vecteurs sont vérifiées.
Le choix de l'amas de points le plus gros est effectué par une méthode similaire à celle utilisée pour la détermination du modèle de couleur. La fonction de poids $g_s^i(u)$, associée au gant i, est alors définie par :

$$\forall \mathbf{u} \in \mathtt{L}^i, g_s^i(\mathbf{u}) = \sum_{\mathbf{x} \in \mathtt{L}^i, \ \mathbf{x} \neq \mathbf{u}} \exp\left[-\lambda_s . d_s(\mathbf{x}, \mathbf{u})\right] \quad (7)$$

La paramètre λ_s permet d'ajuster le calcul du poids. La distance d_s utilisée est une distance de Mahalanobis définie par :

$$d_s(\mathbf{u}_1, \mathbf{u}_2) = [\mathbf{u}_1 - \mathbf{u}_2]^T . \mathtt{C}_s^{-1} . [\mathbf{u}_1 - \mathbf{u}_2] \quad (8)$$

avec $\mathtt{C}_s$, matrice de covariance associée aux 2 composantes spatiales de l'image :

$$\mathtt{C}_s = \begin{pmatrix} \sigma_x^2 & 0 \\ 0 & \sigma_y^2 \end{pmatrix} \quad (9)$$

Le choix de σ_x et σ_y permet d'ajuster la taille des amas que l'on veut détecter. La position du centre de gravité de chaque zone $\hat{\mathbf{u}}^i$ est alors estimée par :

$$\hat{\mathbf{u}}^i = \text{argmax}(g_s^i) \quad (10)$$

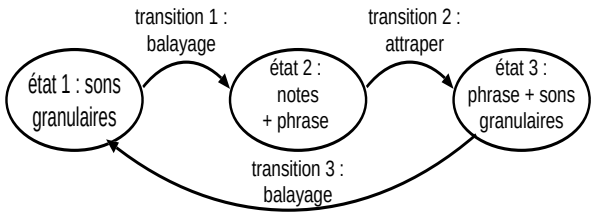

FIG. 3 – Modélisation du système d'interaction geste-son par machine d'état

2.2 Analyse et reconnaissance des gestes

Dans ce module, la trajectoire des poignets est analysée afin de reconnaître plusieurs types de gestes : prendre un objet, lâcher un objet, attraper le son, jouer une note, balayage horizontal.

Prendre un objet, lâcher un objet Le geste de prendre un objet (objet virtuel, représentant soit la phrase, soit un des sons formés) est reconnu par la présence simultanée des deux mains dans son voisinage. L'objet est lâché lorsque les deux mains s'écartent. Il reste alors à la place où il a été lâché.

Attraper le son Le geste d'attraper le son suggère une suspension du temps. Ce geste est reconnu lorsque la distance séparant les mains dépasse un seuil maximum, puis passe sous un autre seuil miminum, dans un petit laps de temps (moins de 0,5 seconde).

Jouer une note L'image est divisée en onze zones verticales. Le geste «jouer une note» est activé lorsqu'un maximum significatif de vitesse verticale est détecté pour l'un des deux poignets.

Balayage horizontal Le geste de «balayage horizontal» suggère l'idée de repousser ou de faire tomber une situation ou un objet. Ce geste est reconnu lorsque la projection sur l'axe horizontal, de la distance entre les deux mains croît rapidement (typiquement : écartement des mains en moins de 0,5 seconde).

2.3 L'interaction geste-son

Le comportement du système d'interaction geste-son peut être modélisé par une machine d'état. Cette dernière est représenté sur la figure 3.
En situation initiale (état 1), la phrase elle-même n'est pas encore jouée, mais deux sons granulaires issus de la phrase sont joués. Un son granulaire est une superposition de plusieurs "grains", fenêtres temporelles de courte durée (quelques 1/10 de secondes), tirées de la même région du fichier audio, ce qui donne l'illusion que le chant s'est figé. Le passage de l'état 1 à l'état 2 est déclenché dès qu'un geste de balayage horizontal est détecté.
Etat 2 : le balayage cause la chute des deux grains. Après un bref silence, la phrase complète s'élève doucement. Dans cette étape, la personne improvise, déclenchant des sons de contrebasse avec le geste "jouer une note".

Lorsqu'un son est attrapé, l'état 3 devient alors actif.

Etat 3 : Le son granulaire attrapé est joué quelques secondes, puis disparaît progressivement. Il est possible d'en attraper plusieurs, successivement. La phrase est jouée, et la contrebasse reste accessible. On termine la séquence en posant les sons au sol, ce qui annule leur volume.

Si on veut reprendre la séquence, un geste de balayage horizontal renvoie à l'étape 1.

Des interactions sonores sont associées à certains gestes :

Attraper le son Lorsque le geste «attraper le son» est détecté, un son granulaire reste figé sur le son joué à cet instant.

Jouer une note Lorsque le geste «jouer une note» est détecté, la bande image dans laquelle ce geste s'est effectué est associée à une des 11 notes de la gamme de la mélodie. La note est alors déclenchée, et son intensité est donnée par la vitesse maximale du geste. Dès que la vitesse passe sous un seuil, on considère le geste terminé, le maximum est effacé, et on est prêt pour le prochain geste.

Comme sur un clavier, la note est tenue tant que le geste ne change pas de sens, c'est à dire tant que la vitesse ne change pas de signe. Ceci constitue un clavier immatériel, qui présente deux caractéristiques intéressantes :

– le déclenchement se fait aussi bien sur un geste vers le haut, que sur un geste vers le bas, ce qui donne de la continuité aux gestes.
– le geste se fait à n'importe quelle hauteur, contrairement à un clavier matériel, où le déclenchement se fait par le contact matériel, donc à hauteur fixe. Ceci ouvre des possibilités (non exploitées ici) : contrôle du timbre par la position verticale du déclenchement ; répartition matricielle des notes...

2.4 Finalités musicales

Deux objectifs, en collaboration avec Gilles Ballet, compositeur, instrumentiste, et enseignant à l'université de Clermont-Fd, ont été visés :

«Matérialisation» du son Tenter de donner au son une dimension matérielle, palpable. Pour ce faire, nous le mettons en relation avec l'espace, par deux procédés :

– Une phrase musicale vocale, mélodique et tonale, est jouée en boucle. Cette phrase est représentée par un objet virtuellement présent dans le champ de la caméra. La personne peut le prendre et le déplacer dans l'espace (en fait dans sa projection plane), ce qui permet de jouer sur deux paramètres (ici, la position verticale modifie le volume, et la position horizontale modifie le panoramique).
– La personne peut à tout instant "saisir" le son avec ses mains. Le son reste alors figé, suspendu, un moment, avant de s'évanouir lentement. La phrase, elle, ne s'arrête pas, ce qui fait apparaître des harmonies. Une fois formé, ce nouvel "objet" peut être saisi et déplacé au même titre que la phrase. C'est une relation entre l'espace et le temps musical.

Déclenchement de notes Il s'agit de donner à la personne la possibilité d'improviser des notes (qui sont des échantillons de contrebasse préalablement enregistrés). Tout geste d'une des mains présentant une composante verticale importante déclenche un son, ce qui met à la disposition de la personne un clavier virtuel à deux dimensions. Celle-ci n'ayant qu'un retour acoustique, on ne peut pas parler de geste instrumental. Néanmoins, la vitesse de chaque geste est prise en compte pour déterminer le volume de chaque note. C'est un jeu d'écoute et d'espace.

3 Réalisation technique

3.1 Traitement du son

Le traitement du son est réalisé sous le logiciel Max/MSP, environnement de programmation graphique par blocs, dédié au traitement en temps réel du son et des événements musicaux au format MIDI.

Les sons de contrebasse ne subissent aucun traitement, autre que la définition de leur volume par la vitesse maxi du geste. Le volume et le panoramique de la phrase sont déterminés par la position de l'objet phrase. Un filtrage passe-bande pourra être utilisé.

La synthèse granulaire utilise 16 grains prélevés dans la même zone de la phrase. Pour donner au son de légères fluctuations, non périodiques, la durée des grains est piochée aléatoirement entre 300 et 400 ms. Ils sont déclenchés toutes les 20ms. Pour éviter les discontinuités, chaque grain est pondéré par une fenêtre de Hanning.

3.2 Acquisition de l'image

Le matériel utilisé pour l'acquisition de l'image est une caméra grand public (ieee1394 ou usb), reliée à un ordinateur portable de type PC. Le traitement du flux vidéo et la détermination des positions des poignets s'opèrent à une cadence d'environ 15 images par seconde. Il convient de noter la présence d'un temps de latence non négligeable, (env. $0.4s$) entre l'acquisition et la restitution du son. Le transfert des informations relatives aux deux positions déterminées vers le logiciel Max/MSP s'effectue via un port MIDI. Cette solution permet de faire tourner le système de vision et la gestion du son sur deux ordinateurs différents.

4 Discussion

Recherche de correspondances geste - fonction musicale Une telle correspondance est pertinente si l'auteur du geste le reconnaît facilement dans le son qui en découle. C'est la condition pour que le système soit lisible, donc pour qu'il y ait une interaction fructueuse avec les gesteurs. La complexité de la correspondance doit être un équilibre entre une correspondance directe, très lisible et vite ennuyeuse, et une correspondance trop complexe, incompréhensible.

Il y a en réalité deux correspondances à établir : correspondance geste-sens, et correspondance sens-fonction musicale. Quel doit être la lisibilité de cette correpondance ? Cela dépend du niveau d'expertise du geste : Un geste ex-

pert peut maîtriser de nombreux paramètres, il attend surtout une stabilité totale du processus qu'il commande. A l'inverse, le système ne sera lisible à un gesteur non expert, que s'il n'a qu'un nombre réduit de paramètres à maîtriser simultanément (1 ou 2).

De ceci découle, qu'un gesteur non expert ne peut interagir qu'avec un système peu variable, donc limité et qui risque de lasser rapidement. Une solution est alors, que sa causalité évolue. Deux voies sont à explorer : évolution séquentielle discontinue (dans laquelle entre le prototype actuel), et apprentissage par la machine, des gestes propres au gesteur présent.

Typologie des gestes à acquérir Une classification des gestes à reconnaître s'impose. L'information gestuelle peut être décrite à trois niveaux :

– bas niveau : coordonnées 2D ou 3D des parties du corps suivies (et leur historique).
– niveau caractérisé : classification d'un geste comme étant : geste périodique, geste direct ou rond, geste statique (posture), gestes corrélés de différentes parties du corps (les deux mains par exemple), geste continu ou impulsif... Des critères appropriés sont évalués pour chaque type de geste : la période pour un geste périodique, l'amplitude et la durée pour un geste direct, la courbure pour un geste rond.
– niveau sémantique : reconnaissance du contenu sémantique des gestes. Ce contenu peut être une action : prendre, modeler, poser, pousser, lancer, rejeter, mettre en tension, relâcher... Ce contenu peut aussi être un sentiment : agitation, sérénité, angoisse, surprise...
Les typologies correspondantes vont être définies en collaboration avec un gesteur professionnel.

5 Conclusion et perspectives

Ce travail présente l'étude et la mise en oeuvre d'un système d'interaction geste-son. Nous avons choisi une approche basée sur le suivi de geste non instrumental par traitement d'image. D'autre types de capteurs sont utilisés, avec des performances élevées, mais leur inconvénient majeur est d'être intrusifs (ils nécessitent que le visiteur porte des capteurs, des fils, ou des marqueurs), et donc d'entraver la spontanéité du geste que nous recherchons. Ceci justifie le choix d'un capteur par vision. Il nous semble pertinent de suivre au moins les mains et le visage, le reste du corps et les membres étant considérés dans un premier temps comme un support et des liens. Par la suite il sera bien sûr intéressant de prendre en compte la totalité de la gestuelle des membres, ainsi que les yeux. Plusieurs points abordés de manière succincte ici mériteraient d'être largement approfondis. Le système présenté ici est un point de départ pour le développement d'interactions gestuelles, avec les activités suivantes, dont le contenu gestuel est riche : danse, mime, marionnette, tai-chi. Nous souhaitons mener des collaborations avec ces disciplines.

Références

[1] C. Cadoz. *Les Nouveaux Gestes de la Musique*. Parenthèses, 1999.

[2] A. Camurri, M. Ricchetti, and R. Trocca. "Eyesweb : Toward Gesture and Affect Recognition in Interactive Dance and Music Systems". *Computer Music Journal*, 24(1) :57–69, 2000.

[3] D. Comaniciu, V. Ramesh, and P. Meer. Real-time tracking of non-rigid objects using mean shift. *Conference on Computer Vision and Pattern Recognition*, 2 :142–149, 2000.

[4] D. Cremers, T. Kohlberger, and C. Schnörr. Nonlinear Shape Statistics in Mumford-Shah Based Segmentation. In *7th European Conference on Computer Vision*, volume 2, pages 93–108, Copenhagen, Denmark, May 2002.

[5] D. Demirdjian and T. Darell. 3-D Articulated Pose Tracking for Untethered Deictic Reference. In *ICMI 2002*, Pittsburgh, Pennsylvania, USA, October 2002.

[6] C. Dobrian and F. Bevilacqua. Gestural Control of Music Using the Vicon 8 Motion Capture System. In *the New Interfaces for Musical Expression 2003 Conference,*, Montréal, Quebec, Canada, 2003.

[7] F. Jurie and M. Dhome. Real time template matching. In *Proc. IEEE International Conference on Computer vision*, pages 544–549, Vancouver, Canada, July 2001.

[8] M. Isard and A. Blake. Condensation – conditional density propagation for visual tracking. *International Journal of Computer Vision*, 29(1) :5–28, 1998.

[9] T. Marrin and J. Paradiso. The digital Baton : a Versatile Performance Instrument. In *International Computer Music Conference*, 1997.

[10] A. Mulder. *Trends in Gestural control of Music*. M.M. Wanderley and M. Battier, IRCAM, Centre Pompidou, 2000.

[11] P. Perez, C. Hue, J. Vermaak, and M. Gangnet. Color-Based Probabilistic Tracking. In *Computer Vision ECCV 2002*, volume 1, pages 661–675, May 2002.

[12] J. Paradis and E. Hu. Interactive Music for Instrumented Dancing Shoe. In *International Music Conference*, 1999.

[13] F. Sparacino. (Some) computer vision based interfaces for interactive art and entertainment installations. *INTER_FACE Body Boundaries*, 2001.

[14] Vicon. Système Vicon. http ://www.vicon.com.

[15] C. Wren and A. Azarbayejani. Pfinder : Real-Time Tracking of the Human Body. In *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1997.