Une étude sur la prise en compte simultanée de deux modalités pour la reconnaissance de gestes de SoundPainting

Irvin Dongo, David Antonio Gómez Jáuregui, Nadine Couture

Une étude sur la prise en compte simultanée de deux modalités pour la reconnaissance de gestes de SoundPainting

A study on the simultaneous consideration of two modalities for the recognition of SoundPainting gestures

Irvin Dongo
Univ. Bordeaux, ESTIA INSTITUTE OF TECHNOLOGY
Bidart, France
Universidad Católica San Pablo
Arequipa, Perú
i.dongoescalante@estia.fr

David Antonio Gómez Jáuregui
Univ. Bordeaux, ESTIA INSTITUTE OF TECHNOLOGY
Bidart, France
d.gomez@estia.fr

Nadine Couture
Univ. Bordeaux, ESTIA INSTITUTE OF TECHNOLOGY,
LaBRI, UMR5800
Bidart, France
n.couture@estia.fr

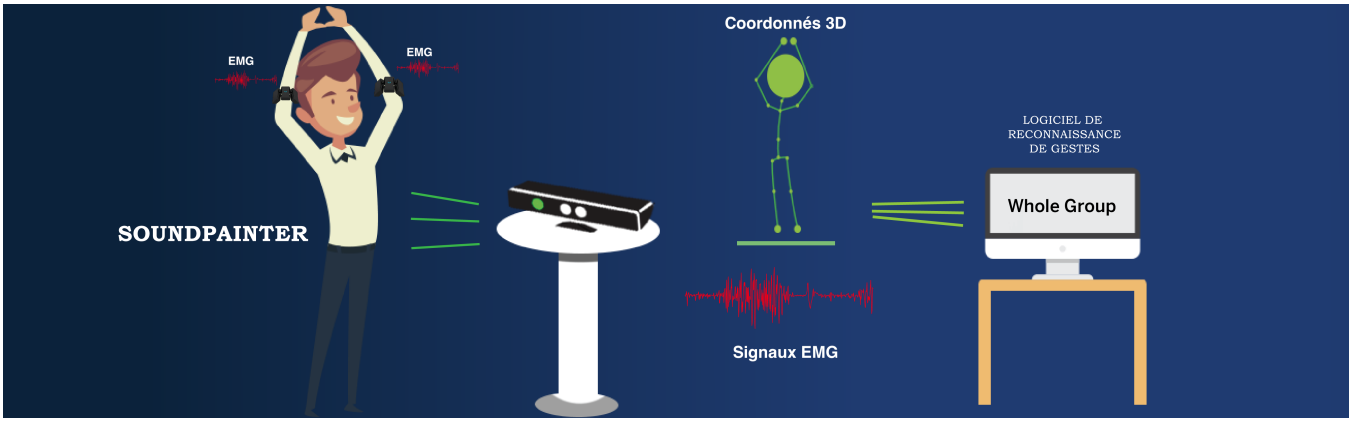

Figure 1: Processus de reconnaissance de gestes, de gauche à droite : le SoundPainter signe un geste de SoundPainting (WholeGroup) ; les données récupérées par les deux capteurs (Kinect® et Myo™) ; caractéristiques extraites des données ; geste reconnu par le modèle de classification.

ABSTRACT

Nowadays, gestures are being adopted as a new modality in the field of Human-Computer Interaction (HMI), where the physical movements of the whole body can perform unlimited actions. Soundpainting is a language of artistic composition used for more than forty years. However, the work on the recognition of SoundPainting gestures is limited and they do not take into account the movements of the fingers and the hand in the gestures which constitute an essential part of SoundPainting. In this context, we conducted a study to explore the combination of 3D postures and muscle activity for the recognition of SoundPainting gestures. In order to carry out this study, we created a SoundPainting database of 17 gestures with data from two sensors (Kinect® and Myo™). We formulated four hypotheses concerning the accuracy of recognition. The results allowed to characterize the best sensor according to the typology of the gesture, to show that a "simple" combination of the two sensors does not necessarily improves the recognition, that a combination of features is not necessarily more efficient than taking

IHM'19, December 10–13, 2019, Grenoble, France

https://doi.org/10.1145/3366550.3372259.

into account a single well chosen feature, finally, that changing the frequency of the data acquisition provided by these sensors does not have a significant impact on the recognition of gestures.

CCS CONCEPTS

• **Human-centered computing** → **Gestural input**; • **Computing methodologies** → *Supervised learning by classification*; *Classification and regression trees.*

KEYWORDS

Gesture recognition, SoundPainting, Machine Learning, Kinect®, Myo™.

RÉSUMÉ

Actuellement, les gestes sont adoptés comme une nouvelle modalité dans le domaine de l'interaction homme-machine, où les mouvements physiques de tout le corps peuvent effectuer des actions quasi-illimitées. Le Soundpainting est un langage de composition artistique utilisé depuis plus de quarante ans. Pourtant, les travaux sur la reconnaissance des gestes SoundPainting sont limités et ils ne prennent pas en compte les mouvements des doigts et de la main dans les gestes qui constituent une partie essentielle de SoundPainting. Dans ce contexte, nous avons réalisé une étude pour explorer la combinaison de postures 3D et de l'activité musculaire pour la reconnaissance des gestes SoundPainting. Pour réaliser cet étude, nous avons créé une base de données SoundPainting de 17 gestes avec les données provenant de deux capteurs (Kinect® et Myo™). Nous avons formulé quatre hypothèses portant sur la précision de la reconnaissance. Les résultats ont permis de caractériser le meilleur capteur en fonction de la typologie du geste, de montrer qu'une "simple" combinaison des deux capteurs n'entraîne pas forcément une amélioration de la reconnaissance, de même une combinaisons de caractéristiques n'est pas forcement plus performante que la prise en compte d'une seule caractéristique bien choisie, enfin, que le changement de la cadence d'acquisition des données fournies par ces capteurs n'a pas un impact significatif sur la reconnaissance des gestes.

MOTS CLÉS

Reconnaissance de geste, SoundPainting, Apprentissage automatique Kinect®, Myo™.

1 INTRODUCTION

L'interaction gestuelle basée sur la vision artificielle est devenue très importante dans l'interaction homme-machine grâce à ses points forts : simplicité, interaction naturelle, intuitive et non intrusive [16]. L'utilisation de l'interaction gestuelle, s'appuyant sur la technologie de la vision artificielle, permet aux artistes de s'exprimer de manière plus libre et moins intrusive par rapport aux interactions gestuelles qui nécessitent des capteurs de mouvement qui peuvent être un peu encombrants et difficiles à étalonner [9]. Le Soundpainting est justement une langue des signes inventée par Walter Thompson en 1974 à Woodstock à destination des artistes. C'est un langage multidisciplinaire et universel qui permet de composer en direct et en temps réel avec des musiciens, des acteurs, des danseurs et des artistes visuels [36, 37]. Ce langage offre une grammaire bien définie, comprenant plus de 1200 gestes, favorisant l'interaction avec un groupe d'artistes dans le cadre de spectacles où les artistes improvisent.

Les travaux sur la reconnaissance des gestes SoundPainting sont encore très limités. A notre connaissance, seuls les travaux de Guyot et Pellegrini [15, 29] et Jáuregui et al.[14] ont exploré la reconnaissance automatique du langage du SoundPainting. Dans ces deux travaux, seule la caméra de profondeur Kinect® est utilisée comme périphérique d'entrée. C'est une limitation technologique car, pour une reconnaissance correcte du geste il est nécessaire de capter également les mouvements des mains et des doigts. Par exemple, des gestes comme *Play/Can't Play*[1], *Walk*[2] and *Continue*[3] sont effectués en utilisant uniquement les mains, ce qui est difficile à capter pour un dispositif caméra en raison de la distance par rapport à l'utilisateur.

Dans [29], les auteurs indiquent qu'il n'existe pas de base de données de SoundPainting dans laquelle des algorithmes d'apprentissage automatique peuvent être exécutés ; c'est pourquoi ils créent une base de données avec 5 répétitions de 20 gestes, mais au final, elle n'est pas partagée avec la communauté scientifique ce qui est une limitation en terme de reproductibilité et de possibilité de tester différents algorithmes de reconnaissance.

Pour surmonter ces deux limitations, nous avons réalisé un étude pour explorer l'utilisation de deux différents capteurs non-invasifs pour la reconnaissance des gestes SoundPainting. Le premier est le capteur Kinect V2®, qui fournit des coordonnées 3D des joints d'articulation du squelette 3D. Le deuxième est le capteur Myo™qui détecte l'activité musculaire de l'avant-bras à partir de signaux EMG (électromyographiques). Ce dernier capteur permet de détecter

[1]Extrait de [37] : *Play/Can't Play* consiste en une série de Cellules (Entrées). Vous (le SounPainter) signez d'abord le geste *Play Can't Play* suivi du nombre de Cellules (Entrées) que vous désirez, ce nombre étant celui que vous voudrez. Le performeur alors en choisit une et une seule et, à l'annonce de sa Cellule, joue sans se soucier de ce que jouent les autres performeurs, comme s'il était seul au monde.

[2]Extrait de [38] : *Walk*, action de marcher. Des deux mains sur un plan horizontal, faire quelques battements à hauteur de la poitrine et un peu en avant du corps.

[3]Extrait de [38] : *Continue*, les performeurs continuent ce qu'ils sont en train d'acter. *Continue* peut aussi se signer d'une seule main en conjonction avec un autre geste, tel que *Point to Point*, ou *Scanning*.

les mouvements de la main initiés par les muscles de l'avant-bras. Ces deux capteurs de faible coût et commerciaux offrent l'avantage de détecter l'activité corporelle et le mouvement des doigts en temps réel sans gêner les artistes lors de leurs performance artistiques. Dans l'étude réalisée, nous avons analysé la contribution des données provenant de ces deux capteurs pour la reconnaissance de gestes SoundPainting en utilisant un algorithme classique de classification (Random Forest [20]). Pour la réalisation de cette étude, nous avons crée une base de données de gestes SoundPainting qui comprend 14 gestes et leurs variations (17 gestes au total). Ces gestes ont été simultanément enregistrés en utilisant une Kinect® et deux capteurs Myo™ qui ont été placés sur chaque avant-bras. Nous avons effectué notre étude grâce à l'implémentation d'un système de reconnaissance de gestes SoundPainting et plusieurs tests de performance. A notre connaissance, aucun système n'a été proposé précédemment pour la reconnaissance des gestes corporels et des mains simultanément à travers des données fournis par ces deux capteurs (Kinect® et Myo™). Le seul système proposé récemment par [12], combine ces deux capteurs pour analyser la corrélation du mouvement de rotation des bras dans une tâche de réhabilitation de mouvement avec un agent virtuel. Toutefois ce système n'a pas été réalisé pour la reconnaissance des gestes. Les deux principales contributions de ce travail sont:

(1) Une base de données du langage SoundPainting qui est mise à disposition[4] pour la communauté des chercheurs.
(2) Une étude sur la combinaison de données de posture 3D et de l'activité musculaire (EMG) pour la reconnaissance de gestes impliquant simultanément des mouvements du corps et des mains.

Nous organisons ce papier comme suit. La section 2 examine la littérature correspondante. La section 3 décrit le processus de reconnaissance de gestes proposé. Dans la section 4, nous détaillons comment la base de données SoundPainting a été créée. La section 5 présente l'implémentation de notre approche. La section 6 décrit les expériences menées afin d'évaluer quantitativement la précision de la reconnaissance des gestes SoundPainting. Les résultats sont discutés dans la section . Enfin, nous concluons dans la section 8.

2 ÉTAT DE L'ART

Reconnaissance des gestes du SoundPainting

Les auteurs de [15, 29] ont développé un système de reconnaissance des gestes SoundPainting basé sur un Kinect® en tant que périphérique d'entrée et des modèles de Markov cachés pour les phases d'apprentissage et de classification. Les auteurs affirment, en 2014, que la proposition est la première reconnaissance de geste appliquée au SoundPainting. Ils proposent également un corpus d'enregistrements vidéo de Christophe Mangou[5], en effectuant environ 20 gestes répétés 5 fois chacun. Cependant, ils ne précisent pas quels gestes ont été enregistrés et il n'y a aucune référence à la base de données pour réutiliser les données par d'autres recherches. De notre point de vue, 5 répétitions par geste ne suffisent pas pour créer un modèle de classification de gestes qui présentent certaines similitudes. En effet, avec 5 répétitions avec une Kinect®, compte tenu des similitudes, peu de points sont obtenus et donc la segmentation du geste obtenue est très peu précise.

Dans [14], très récemment en 2019, les auteurs ont développé un système de reconnaissance des gestes de SoundPainting pour générer des sons de musique électronique. L'originalité du système est de permettre au SoundPainter de s'adapter très rapidement au style (c'est à dire à la façon de signer) des gestes qui ont été appris. C'est une contrainte, car il faut que le SoundPainter signe "dans le style appris", celui du créateur du SoundPainting, Walter Thomson. Mais, c'est aussi un avantage, car dans un contexte pédagogique cela entraîne les apprentis SoundPainter à signer correctement. Cette proposition est basée sur une pose clé pour chaque geste et sur un arbre de décision pour créer un modèle de classification. Toutefois, la sélection de la pose clé est manuelle, seules 6 postures sont reconnues et la temporalité du geste n'est pas pris en compte.

Reconnaissance des gestes réalisés avec tout le corps

Plusieurs travaux [7, 17, 23] utilisent les images de profondeur fournies par ces caméras car elles ne sont pas impactées par les changements brusques d'éclairage dans des environnements non contrôlés. En outre, la disponibilité des coordonnées 3D des articulations extraites à partir de ces images permet d'avoir une description précise de la configuration du corps humain [8]. La plupart des approches de reconnaissance de gestes utilisent des algorithmes d'apprentissage automatique qui s'entraînent et apprennent à partir de caractéristiques précises de la posture du corps humain [8, 41]. Deux des algorithmes les plus utilisés pour la reconnaissance gestuelle sont les machines à vecteurs supports [23] et les modèles de Markov cachés [40]. En 2017, les méthodes d'apprentissage profond (*deep learning*) ont permis de grandes avancées pour le problème de reconnaissance gestuelle [5]. Toutefois, cette dernière approche est difficile à mettre en œuvre en raison de la très grande quantité de données nécessaire pour entraîner ce type de modèle.

[4] https://www.dropbox.com/sh/jrzmj14wm3icdbs/AAB4DaAjuE-zd8U7fG3rJgy9a?dl=0

[5] Chef d'orchestre de formation classique, attiré par tous les genres d'expression musicale, du répertoire classique à l'improvisation en passant par le jazz, le rock et le Soundpainting - https://christophemangou.com

Reconnaissance des gestes des mains

En raison de la particularité des gestes de SoundPainting, où les mains et les doigts jouent un rôle important, les travaux sur le suivi et la reconnaissance des gestes de la main sont analysés. Dans [39], les auteurs utilisent un gant multicolore pour reconstituer la pose de la main à partir d'une seule image. D'autres propositions sont basées sur des caméras de profondeur [27, 31]. Par exemple, les auteurs dans [27] utilisent un modèle 3D de la main généré à partir d'une image couleur RGB d'une main et sur l'image de profondeur correspondante fournies par le Kinect®. Dans [31], les auteurs affirment que le suivi des mains est difficile, car les mains peuvent former diverses postures complexes en raison de leurs nombreux degrés de liberté (DoF), et se présentent sous différentes formes et tailles. Ils utilisent une caméra de profondeur pour leur système avec un "réinitialiseur" par image qui assure une récupération robuste en cas de perte de trace. A partir des algorithmes de suivi de mains, des algorithmes de reconnaissance gestuelle peuvent être utilisé pour la reconnaissance de la langue de signes. Nikam et al.[25] ont proposé un système en temps réel pour la reconnaissance des gestes de la main sur la base de la détection (par traitement d'image) de certaines caractéristiques et propriété géométriques (l'orientation, le centre de la masse, la position des doigts et le pouce). Shahriar et al. [30] ont développé un système de traduction dactylologique ASL (la langue des Signes Américaine) à partir d'un algorithme de segmentation de la peau avec un apprentissage profond (deep learning).

Plusieurs dispositifs commerciaux de suivi et de reconnaissance des gestes des mains sont devenus récemment disponibles sur le marché. Par exemple, *CyberGlove III* qui comporte de 18 à 22 capteurs, offre une grande précision et une faible linéarité du capteur (moins de 0,6%). Leap Motion [2], qui est basé sur une caméra de profondeur, est capable de suivre les deux mains en même temps et est devenu très populaire et très largement utilisé. Cependant, la plage idéale de capture du contrôleur Leap Motion s'étend d'environ 25 à 600 millimètres au-dessus du périphérique; Cette technologie n'est donc pas adaptée à la reconnaissance des mains du SoudPainter, car elle limiterait sa performance en contraignant trop son espace d'expression. La société North®a développé un bracelet, appelé Myo™, basé sur des capteurs électromyographie (EMG) pour détecter l'activité musculaire. Il possède également des capteurs de type gyroscope, accéléromètre et magnétomètre. Par défaut, 5 gestes sont reconnus, mais par l'intermédiaire de son SDK, de nouveaux gestes peuvent être appris. Cet appareil est utilisé dans plusieurs travaux relativement récents comme [10, 11] pour la performance musicale (violonistes), pour la classification du mouvement des doigts par [32], pour l'analyse de la navigation par geste de la main par [24], pour la réalité virtuelle par [22], pour la création d'un mapping interactif entre les gestes et des sons musicaux [34] etc. Le Myo™ offre plusieurs avantages par rapport aux autres appareils : un coût abordable (environ 250 €), une configuration simple, et la possibilité d'être caché ce qui est très intéressant en terme de performances artistiques [35].

3 LES CINQ PHASES DU PROCESSUS DE RECONNAISSANCE

Afin de réaliser notre étude, nous avons développé un système de reconnaissance des gestes qui comprend cinq phases (voir Figure 2) : (i) une *phase de synchronisation* dans laquelle les données des deux capteurs (Kinect® et Myo™) sont acquises de façon synchrone à la même cadence ; (ii) une *phase de segmentation de gestes*, où un mouvement est divisé en plusieurs gestes ; (iii) une *phase d'extraction de caractéristiques* dans laquelle les données obtenues par les périphériques d'entrée sont traitées pour obtenir les meilleurs descripteurs ou caractéristiques des gestes ; (iv) une *phase d'apprentissage* qui consiste à construire un modèle de classification des gestes en utilisant des techniques d'apprentissage automatique ; et enfin, (v) une *phase de classification*, où les gestes sont reconnus par le modèle de classification appris au cours de la phase précédente. Ces cinq phases sont décrites dans les paragraphes suivants.

Phase de synchronisation

Lorsque les périphériques d'entrée ont des cadences d'acquisition différents, ce qui est notre cas, une phase de synchronisation est nécessaire pour obtenir les données dans la même cadence. Dans le cas de notre système, la cadence d'acquisition du Kinect® est de 30Hz et celui du Myo™ est de 200Hz. Notre phase de synchronisation consiste à définir la même cadence d'acquisition pour la capture des données. Nous avons testé plusieurs cadences. Par exemple, pour une cadence de 30Hz, qui est celle du capteur Kinect, la cadence d'acquisition du capteur Myo a été réduite à 30Hz en utilisant la méthode de décimation du filtre FIR (Finite Impulse Response) [3]. Pour une cadence à 200Hz, celle du capteur Myo, la cadence du capteur Kinect a été augmenté à 200Hz en utilisant une méthode d'interpolation basé sur l'échantillonnage de Nyquist [1].

Phase de segmentation

Un mouvement peut contenir un ou plusieurs gestes qui peuvent également être composés de plusieurs sous-gestes. Ainsi, le langage SoundPainting comporte des gestes plus ou moins complexes à interpréter. Par exemple, *Whole Group*[6]

[6]Extrait de [36] : *Whole Group* signifie tous les performeurs - l'ensemble entier. Mettre les 2 bras au dessus de votre tête, créant un cercle, le bout des doigts se rejoignant à peine.

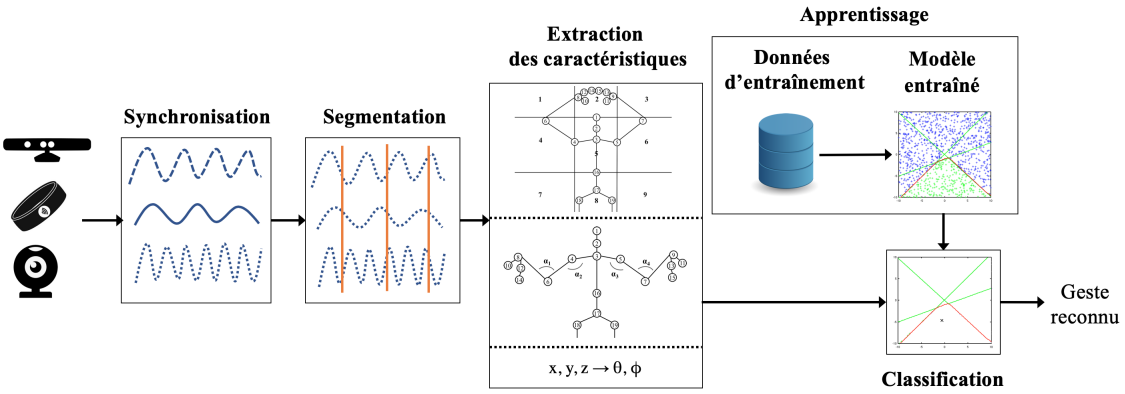

Figure 2: Les phases de notre système

(visible sur les figures 1 et 4) a une seule manière d'être exécuté, alors que *Scanning*[7] peut être signé par un bras, ou l'autre, ou bien les 2 en même temps. Nous considérons un geste, noté g, comme une séquence de poses clés : $g = [kp_1, kp_2, ..., kp_n] \mid \forall kp_i, kp_i$ est une pose clé. Afin de fournir une segmentation appropriée, nous proposons l'utilisation d'une pose clé d'initialisation (les bras se trouvent le long du corps) (kp_{init}) pour activer le mouvement. Ainsi, un geste de SoundPainting, noté g_{SP}, est délimité par une pose clé initiale : $g_{SP} = [kp_{init}, g_1, g_2, ..., g_n, kp_{init}] \mid \forall g_i \in G$ où kp_{init} est la pose clé initialisation et où G est l'ensemble de tous les sous-gestes. Par exemple, si g_W est le geste *Whole Group*, $g_W = [kp_{init}, g_1, g_2, kp_{init}]$ où g_1 est le mouvement emmenant les bras au dessus de la tête et g_2 est le mouvement remmenant les bras le long du corps (pose clé initiale). Pour améliorer l'efficacité du système, nous prenons en compte les "pauses" (temps de pause γ) comme mesures permettant ainsi de détecter les gestes au fur et à mesure de leur exécution (dans le flux). Par exemple, entre g_1 et g_2, une pause s'opère, la reconnaissance du geste *Whole Group* est réalisé. Lorsque kp_{init} est reconnu, le geste est fini et le système est prêt à commencer un nouveau geste.

Phase d'extraction de caractéristiques

Selon le type de périphérique, certaines situations d'usage peuvent affecter les données brutes fournies par les capteurs, qui pourtant devraient être similaires pour les mêmes gestes. Par exemple, pour les périphériques d'entrée basés sur des caméras RGB et profondeur (comme c'est le cas avec la Kinect®), différentes distances et positions du corps par rapport à la caméra, fournissent des valeurs brutes différentes. De plus, la morphologie et la taille du corps humain ont un impact sur les données fournies par les périphériques. Ainsi, notre phase d'extraction de caractéristiques consiste en un ensemble de fonctions qui calculent des nouvelles valeurs à partir des données reçues des capteurs. Ces nouvelles valeurs sont appelées *caractéristiques* et sont utilisées, dans notre cas, pour mieux décrire les poses clés. Par conséquent, un ensemble des poses clés est formalisée en tant que vecteur $kp_{d_i} = [f_1(DA_{d_1}), f_2(DA_{d_2}), ..., f_n(DA_{d_n})]$, où DA correspond aux données obtenues du capteur d_i et f_i la fonction d'extraction de caractéristiques. Selon nos hypothèses, les deux capteurs peuvent être utilisés pour reconnaître un geste, dans ce cas, une pose clé pour le système est l'union de toutes les caractéristiques extraites de chaque capteur, $kp_s = kp_{d_1} \bigcup kp_{d_2}$.

Phase d'apprentissage

Une fois que les caractéristiques décrivant une pose clé sont obtenues lors de la phase d'extraction des caractéristiques, nous effectuons un processus d'apprentissage à l'aide de techniques d'apprentissage automatique. Selon la définition d'un geste g, il est composé d'une séquence de poses clés kp, dont la taille varie selon chaque geste. Afin de fournir une entrée adéquate pour les techniques d'apprentissage automatique, nous appliquons un processus de réduction de toutes les poses appartenant à un geste à partir des techniques de classification (clustering, avec l'algorithme du K-means [21]). Ainsi, nous avons sélectionné les poses clés les plus pertinentes permettant de décrire chaque geste. Nous réduisons l'ensemble des N poses à seulement n poses clés, où n est un paramètre défini arbitrairement tel que $n < N$. Ainsi, un geste est redéfini comme un vecteur $g = [kp_1, kp_2, ..., kp_n]$, où kp_i est une pose clé sélectionnée par les techniques de classification. Il est conseillé, pour simplifier la réduction, de prendre n comme nombre de classes (clusters) de la méthode de réduction et de prendre les poses de clé comme le centroïde de chacune des classes (clusters). Un échantillon de plusieurs répétitions pour chaque geste

[7] Extrait de [36] : *Scanning* Les performeurs répondent par une improvisation libre lorsque le geste de balayage passe devant eux, qu'elle qu'en soit la direction. Les performeurs cessent immédiatement de jouer dès que le geste du Balayage n'est plus en face d'eux. Doigts serrés et paume tournée vers l'ensemble, tendez votre bras d'un coté, quelques centimètres au-dessus de votre épaule. Gardant le bras tendu, faites le geste de le faire passer par dessus les têtes de l'ensemble, dans un mouvement de Balayage.

est nécessaire afin de créer un bon modèle de classification. L'entrée des techniques d'apprentissage automatique est composée de deux vecteurs : le premier est le vecteur des poses clés de gestes et le second leurs classes respectives: $< \{g_1, g_2, ..., g_n\} >, < \{class_{g_1}, class_{g_2}, ..., class_{g_n}\} >$.

Phase de classification

Cette phase de classification dépend fortement de la réduction de des poses appartenant à un geste SoundPainting à partir des techniques de classification décrite dans la phase d'apprentissage). Elle consiste en la prédiction du geste reçu, autrement dit l'ensemble des poses résultantes de la réduction, afin de classer en temps réel le geste effectué par la personne qui signe (pour nous, le SoundPainter). Le modèle prédictif détermine donc la classe (c'est à dire le geste SoundPainting) avec la probabilité la plus élevée à laquelle appartient le vecteur d'entrée (ensemble des poses clés d'un geste SoundPainting). Le résultat de la phase de classification est donc un ensemble de probabilités de toutes les classes (c'est à dire tous les gestes SoundPainting) apprises dans la phase d'apprentissage. Avec cette méthode, il faut noter que le système reconnaîtra toujours un geste comme appartenant à notre base de données.

4 BASE DE DONNÉES SOUNDPAINTING

Pour choisir les gestes de SoundPainting à enregistrer dans la base, nous avons testé avec un Kinect® et avec la collaboration d'un SoundPainter expérimenté les 30 gestes que ce SoundPainter utilise lors de ses performances. Nous avons constaté que le squelette fourni par le Kinect® est faux lorsque les gestes utilisent les bras croisés. Nous avons donc sur les 30 gestes, conservé seulement 14 gestes. Dans le cas de Myos™, les données reçue sont toujours correctes. Pour capturer les données, afin d'alimenter la base de données, trois participants/opérateurs ont contribué et nous avons utilisé : 4 ordinateurs Intel Core(TM) i5 à 2,6 GHz avec 8,00 Go, une caméra RGB (résolution 1280x1024, 30Hz), deux capteurs Myos™, une caméra de profondeur Kinect®. Chaque participant, à tour de rôle, a réalisé 50 répétitions des 14 gestes de Soundpainting choisis, dont 3 gestes avec 2 variations, faisant un total de 17 gestes. Dans notre base de données, les données de chaque capteur sont séparées dans des fichiers différents. De cette manière un utilisateur pourra choisir le capteur à utiliser. Chaque geste Soundpainting a été réalisé à partir d'une posture d'initialisation (les deux bras le long du corps). Les 14 gestes Soundpainting sont décrits dans le tableau 1.

Procédure

Le capteur Kinect® est placée en face du participant équipé des deux capteurs Myos™ (un capteur Myo™ pour chaque avant-bras). Les deux Myos sont orientés dans le même sens.

Table 1: Nom du geste Soundpainting (les gestes sont décrits dans les références suivantes: [36], [37], [38])

Numéro du geste	Nom du geste
1	Whole Group
2	Rest of the Group
3	Long Tone
4	Pointillism
5	Hit
6	Silence
7	Minimalism
8	Scanning (2 variations)
9	Play
10	More space fader
11	Off (2 variations)
12	Match (2 variations)
13	Finish your idea
14	Wait

La distance entre le participant et le capteur Kinect® a été de 3m20 afin de pouvoir enregistrer tout le corps du participant. Le système de capture des données a nécessité 4 ordinateurs. Un ordinateur héberge l'interface de contrôle pour lancer l'enregistrement des données des 3 capteurs. Deux ordinateurs récupèrent les informations issues des deux Myos™. Le quatrième ordinateur récupère les données du capteur Kinect®. Le capteur Kinect® est capable de fournir les coordonnées 3D de 25 joints des articulations du corps entier. Le capteur Myo™ est capable de détecter et quantifier 8 signaux électriques provenant des contractions musculaires de l'avant bras. Pour chaque geste, une capture vidéo a été prise avec une caméra classique afin de vérifier que le geste a été bien réalisé. La figure 3 montre l'installation du système d'enregistrement des données.

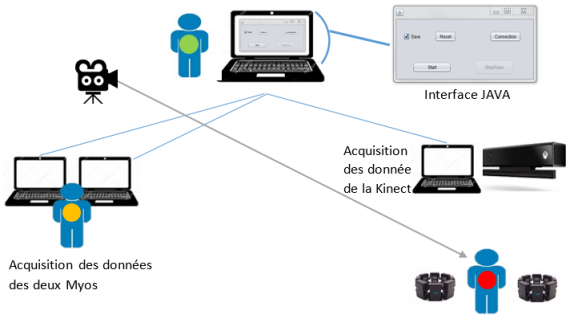

Figure 3: Installation du système d'enregistrement des données de gestes

5 IMPLÉMENTATION DES PHASES

Pour réaliser notre étude, un prototype a été développé à l'aide de TensorFlow[8] et Python. Dans le but de ne pas obliger la personne à rester immobile face au Kinect®, nous

[8]TensorFlow est une plate-forme open source d'apprentissage automatique. Il offre un écosystème complet et flexible d'outils, de bibliothèques et de ressources communautaires qui permet aux chercheurs d'avancer dans l'apprentissage automatique et aux développeurs de créer et de déployer facilement des applications. https://www.tensorflow.org

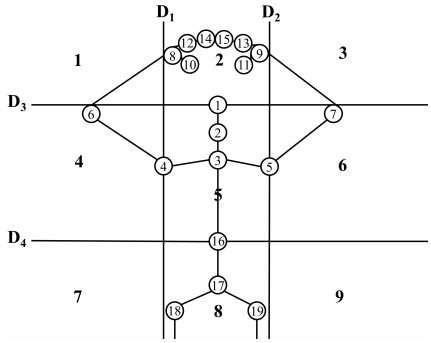

Figure 4: Caractéristiques indépendantes : positions des mains

avons fait, dans notre phase d'extraction des caractéristiques, un changement de repère, en passant du repère absolu du Kinect® à un repère relatif dont l'origine a été placée au point d'articulation 16, nommé "Colonne Vertébrale" (voir figure 4). Cela permet donc au compositeur (SoundPainter) de pouvoir se déplacer, en veillant cependant à rester dans le champ de vision du Kinect®. Nous avons extrait trois caractéristiques du Kinect® (les positions des mains, les angles des bras et les points d'articulations passés en coordonnées sphériques) et deux caractéristiques du capteur Myo™ (valeur absolue moyenne et longueur de la forme d'onde). Cet ensemble de caractéristiques est décrit dans les sous-sections suivantes.

Première caractéristique extraite du Kinect® (FK1) : Positions des Mains. Les mains sont les parties du corps qui contiennent le plus de variations. Prendre leurs positions comme caractéristiques est donc indispensable. Très utilisée en reconnaissance gestuelle, nous avons utilisée la méthode [6] qui consiste à diviser l'espace autour du corps en 9 cadrans définis par 4 droites (D_1, D_2, D_3 et D_4) comme indiqué sur la figure 4. Dans l'exemple de la figure 4, lorsque le geste *Whole Group* est fait, la main droite et la main gauche sont dans le cadran 2 car les articulations 12 et 13 correspondantes le sont. Les positions des mains sont formalisées par l'équation 1.

$$PM(J) = [P(j_8), P(j_9)] \tag{1}$$

où $P(j_i)$ est la fonction qui retourne le cadran du point d'articulation j_i.

Deuxième caractéristique extraite du Kinect® (FK2) : Angles des bras. Les quatre angles générés par les bras, représentés sur la figure 5, sont souvent utilisés comme des caractéristiques car ils sont invariants aux changements d'échelle et à la translation [41], donc indépendants de la morphologie. De plus, ils complètent l'information de la position des mains. Les valeurs des angles sont formalisées par l'équation 2.

$$AB(J) = [A(j_8, j_6, j_4), A(j_6, j_4, j_3), A(j_3, j_5, j_7), A(j_5, j_7, j_9)] \tag{2}$$

où $A(j_i, j_j, j_k)$ est la fonction qui calcule l'angle entre les points d'articulation j_i, j_j et j_k.

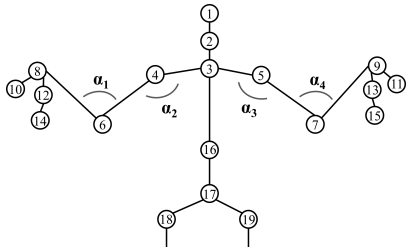

Figure 5: Caractéristiques indépendantes : 4 angles des bras

Troisième caractéristique extraite du Kinect® (FK3) : Coordonnées sphériques. Selon [33], un moyen de simplifier le processus de normalisation pour rendre le squelette indépendant de la morphologie et de la taille, est de convertir les coordonnées cartésiennes des points d'articulation en coordonnées sphériques. La conversion se calcule ainsi :

$$\rho = \sqrt{x^2 + y^2 + z^2} \qquad \theta = \arccos\frac{z}{r} \qquad \phi = \arctan\frac{y}{x}$$

Pour nous affranchir de la morphologie des personnes, nous considérons θ et ϕ comme des caractéristiques. Nous ne prenons pas ρ comme une caractéristique car c'est une distance (distance radiale) qui dépend de la morphologie.

Dans les 14 gestes à reconnaître les jambes n'interviennent pas. Donc, pour optimiser la reconnaissances, nous ne considérons pas les 8 articulations numérotées de 17 à 25 et ne conservons que les articulations de 1 à 16. Avec θ et ϕ, un vecteur de taille 32 est généré (équation 3).

$$CS(J) = [\theta_{j_1}, \phi_{j_1}, \theta_{j_2}, \phi_{j_2}, ..., \theta_{j_{17}}, \phi_{j_{16}}] \tag{3}$$

où θ_{j_i} et ϕ_{j_i} sont les coordonnées sphériques du point d'articulation j_i.

Première caractéristique extraite du Myo™ (FM1) : Valeur absolue moyenne. MAV (*Mean Absolute Value*) est l'estimation de la valeur absolue de la somme ; Elle mesure le niveau de contraction des signaux EMG [4]. Elle est définie comme suit :

$$MAV(E) = \frac{1}{N}\sum_{k=1}^{N}|e_k| \tag{4}$$

où e_k est le signal EMG à k (un nombre entier qui varie de -128 à 128) et N est le nombre d'échantillons (c'est à dire le nombre de poses clés). Parce que le capteur Myo™ a 8 signaux EMG, sa valeur absolue moyenne est calculée comme un vecteur de 8 valeurs :

$$MAV(Myo) = [MAV(E_1), MAV(E_2), ..., MAV(E_8)] \tag{5}$$

où $MAV(E_i)$ est le valeur absolue moyenne du signal E_i.

Deuxième caractéristique extraite du Myo™ (FM2) : Longueur de la forme d'onde. WL (*Waveform Length*) est une variation cumulative pouvant indiquer les variations de degré des signaux EMG [4]. Il est défini comme suit:

$$WL(E) = \sum_{k=1}^{N} |e_{k+1} - e_k| \tag{6}$$

Comme MAV, la longueur de la forme d'onde du capteur Myo™ est composé de 8 valeurs.

$$WL(Myo) = [WL(E_1), WL(E_2), ..., WL(E_8)] \tag{7}$$

où $WL(E_i)$ est la longueur de la forme d'onde du signal E_i.

Implémentation des phases. Pour la segmentation des gestes, nous avons utilisé un temps de pause de 840 ms (γ) qui représente le temps minimum entre la fin d'un sous-geste et le début du suivant. Pour les phases d'apprentissage et de classification, nous avons utilisé l'algorithme *Random Forest[20]* [9] qui est très utilisé dans la littérature de la reconnaissance des gestes [13, 26]. Pour la sous-phase de réduction des poses clés, une méthode *K-means [21]* a été utilisé afin de réduire un geste à seulement 5 poses clés pertinents. Cette valeur de 5 a été trouvé expérimentalement afin d'obtenir le meilleur compromis entre le temps de calcul de la réduction des poses et la précision de l'algorithme classifeur.

6 ÉVALUATION EXPÉRIMENTALE

Le système, décrit ci-dessus, a été utilisé pour effectuer plusieurs tests afin d'évaluer la précision obtenue par les données brutes et les caractéristiques extraites de chaque capteur et les combinaisons de ces caractéristiques. Nous avons évalué également la précision par rapport à la différence de cadence d'acquisition de chaque capteur et le temps de calcul obtenu pour chacune des phases du système. Les hypothèses de notre étude sont les suivantes :

H1 : La précision obtenue à partir du capteur Kinect® est supérieure à celle du capteur Myo™ lorsque les gestes SoundPainting impliquent plus des mouvements corporels que des mouvement des mains.

H2 : La précision obtenue à partir de la combinaison des deux capteurs est supérieure à la précision obtenue à partir d'un seul capteur (Kinect® ou Myo™).

H3 : Pour chaque capteur, la précision obtenue à partir de la combinaison des caractéristiques est supérieure à l'utilisation d'une seule caractéristique.

H4 : Le changement de la cadence d'acquisition des capteurs Kinect® et Myo™ n'a pas un impact significatif sur la précision du système.

Configuration

Pour réaliser notre étude, nous avons considéré deux séquences contenant 116 gestes en total. Chaque séquence est exécuté par une personne différente (2 hommes, 28 ans et 38 ans) se positionnant à des distances et des positions aléatoires vis à vis de la Kinect®, mais néanmoins dans son champ de vision. Cette configuration peut être appliqué facilement dans une performance en live car le SoundPainter doit rester toujours en face des musiciens sans changer sa position dans le scénario[10]. La durée de la première séquence de gestes a été de 3 minutes 4 secondes et de la deuxième séquence a été de 6 minutes 35 secondes. Le logiciel de reconnaissance a été exécuté sur un MacBook Pro, Intel Core(TM) i7 à 2,2 GHz avec 16,00 Go, exécutant un MacOS Mojave et utilisant un environnement de programmation Python 2.7.10 et TensorFlow version 1.12.0.

Mesure de la précision

Pour évaluer la précision des deux modalités et leur combinaison, nous avons utilisé la formule de précision suivante:

$$\mathbf{PR} = \frac{VP}{VP + FP} \in [0, 1] \tag{8}$$

où VP (Vrais Positifs) est le nombre de gestes correctement reconnus et FP (Faux Positifs) est le nombre de gestes mal reconnus.

Évaluation en 3 tests

Nous avons réalisé 3 différents tests pour évaluer expérimentalement la contribution des deux modalités. Dans le test n°1, nous avons évalué la précision (PR) de notre système obtenue par chaque caractéristique extraite de chaque capteur (Kinect® et les capteurs Myo™) ainsi que les combinaisons de toutes les caractéristiques. Pour le test n°2, nous avons évalué la performance de notre système avec les deux cadences d'acquisition de chaque capteur (30 Hz et 200 Hz). Finalement, dans le test n°3, nous avons évalué le temps de calcul du système proposé pour les phases d'extraction de caractéristiques, la phase de réduction et la phase de classification.

Test n°1 - Précision de la reconnaissance des gestes : Les tableaux 2, 3 et 4 présentent la précision de la reconnaissance des gestes obtenue par chaque caractéristique implémentée dans notre système ainsi que ses combinaisons : les positions des mains $FK1$, les angles des bras $FK2$, les coordonnées sphériques $FK3$, la valeur absolue moyenne des signaux

[9] Les forêts aléatoires (*Random Forest*) sont une méthode d'apprentissage automatique qui consiste à créer une série d'arbres de décision au moment de la formation et à retourné la classe avec le vote le plus élevé de tous les arbres de l'ensemble - http://www.nickgillian.com/wiki/pmwiki.php/GRT/RandomForests.

[10] https://christophemangou.com

EMG *FM*1 et la longueur de la forme d'onde des signaux EMG *FM*2. La précision de la reconnaissance a été calculée pour chaque geste Soundpainting appris et enregistré dans la base de données.

Test n°2 - Précision par rapport à la cadence d'acquisition : Le tableau 5 présente la précision (PR) de la reconnaissance gestuelle obtenue en utilisant les deux différentes cadences d'acquisition obtenues par chaque capteur. Les caractéristiques extraites pour ce test sont celles qui ont obtenu les meilleurs résultats de performance (pour chaque capteur) lors du test n°1.

Test n°3 - Temps de calcul : Le tableau 6 présente le temps de calcul (milisecondes) de la phase d'extraction des caractéristiques, de la phase de classification (réduction avec K-means et prédiction) et le temps de calcul total. Tous ces temps de calcul sont présentés pour les combinaisons de caractéristiques. Le temps de calcul présenté correspond au temps de calcul moyen de 5 exécutions du système avec les 2 séquences vidéo (120 gestes).

7 DISCUSSION

Les résultats (précision moyenne) des tableaux 2 et 3 montrent que la précision de la reconnaissance des gestes obtenue par le capteur Kinect® est globalement supérieur (FK2+FK2+FK3 = 48,49%) à la précision obtenue par le capteur Myo™ (FM1+FM2 = 21,87%). Ces résultats suggèrent que dans la plupart des 17 gestes SoundPainting le mouvement corporel est plus discriminant que le mouvement des mains ou des doigts. Toutefois, pour les gestes *Hit* (FM1+FM2 = 42,86%, FK1+FK2+FK3 = 0%), *Scanning Right* (FM1+FM2 = 50%, FK1+FK2+FK3 = 45,45%), *Scanning Left* (FM1+FM2 = 100%, FK1+FK2+FK3 = 75%) et *Match Left* (FM1+FM2 = 50%, FK1+FK2+FK3 = 0%) la précision de la reconnaissance par le capteur Myo™ a été supérieur à celle du capteur Kinect®. Dans ce cas, les résultats suggèrent que les mouvements des mains ou des doigts sont plus variés. Ces résultats vérifent l'hypothèse **H1**.

Les résultats (précision moyenne) du tableau 4 montrent que la combinaison de toutes les caractéristiques de deux capteurs n'améliorent pas globalement la précision de chaque capteur utilisé séparément. Ces résultats rejettent l'hypothèse **H2**. Toutefois les résultats ont montré deux cas dans lesquels la combinaison des deux capteurs ont amélioré la précision. Ces deux cas sont présentes pour les gestes *Hit* (FK3 = 0%, FM1+FM2 = 42,86%, FK3+FM1+FM2=50%) et *Off Right* (K_{Brutes} = 0%, M_{Brutes} = 0%, K_{Brutes}+M_{Brutes} = 50%). Ces résultats suggèrent également qu'une combinaison simple à partir de l'union de ces caractéristiques peut détériorer la performance du classifieur. Afin de résoudre ce problème, des méthodes plus adaptées de fusion des données multimodales [18] ou de fusion de classifieurs [19] peuvent être intégrés pour améliorer la précision de la reconnaissance des gestes à partir des deux capteurs.

Les résultats des tableaux 2 et 3 montrent que la précision obtenue par la combinaison de plusieurs caractéristiques provenant du même capteur n'est pas toujours supérieure à la précision obtenue par une seule caractéristique extraite (par exemple pour le geste *Pointillism*: FK1+FK2 = 33,33%, FK2 = 66,67%). Ici, afin d'améliorer la précision, des analyses de corrélation seront nécessaires pour sélectionner uniquement les caractéristiques qui peuvent être combinées dans le modèle de classification. Ces résultats rejettent l'hypothèse **H3**.

Les résultats du tableau 5 montre un écart de précision selon la cadence utilisée pour les capteurs, 30 Hz ou 200 Hz. Toutefois cette différence est de 3,37%. Ce résultat indique que le changement de la cadence n'a pas un impact significatif sur la précision de la reconnaissance des gestes. L'hypothèse **H4** est donc acceptée. Finalement, les temps de calculs reportés dans le tableau 6 permettent de constater que environ 80% du temps d'exécution est occupé par la réduction des poses clés. Ce temps de calcul pourrait être réduit en utilisant un algorithme de type K-means bien adapté pour des calculs en temps-réel [28].

8 CONCLUSION

Dans cet article nous avons réalisé un étude pour explorer l'utilisation de deux modalités, posture 3D et activité musculaire des avant bras, pour la reconnaissance des gestes SoundPainting. Les données sont extraites à partir de deux capteurs (Kinect® et Myo™). Pour réaliser cet étude, nous avons crée une base de données des gestes SounPainting et nous avons implémenté un système qui combine les données de posture 3D et les données de l'activité musculaire des avant bras pour classifier des gestes SoundPainting. Des analyses de précision ont été effectué pour étudier la contribution de chaque caractéristique extraite des capteurs et de leur combinaison par rapport à la précision de la reconnaissance gestuelle. Les résultats de précision même s'ils sont relativement faibles, suggèrent qu'une combinaison des deux capteurs est capable d'améliorer la reconnaissance des gestes SoundPainting principalement pour les gestes impliquant plus des mouvements des mains ou des doigts. Pour aller plus loin, des nouvelles études de corrélation devront être effectuées ainsi que l'implémentation de nouvelles approches de fusion de ces données multimodales [18] ou de fusion de classifieurs [19]. Ceci n'est pas simple du fait de la nature très différentes de ces données.

Nous travaillons actuellement à étendre ce travail en extrayant de nouvelles caractéristiques à partir de ces deux capteurs. Nous envisageons également d'améliorer la précision en utilisant des techniques d'apprentissage automatique plus performantes comme l'apprentissage profond (Deep

Table 2: Précision obtenue par les caractéristiques extraites à partir du capteur Kinect

Geste	Précision							
	Brutes	FK1	FK2	FK3	FK1+FK2	FK1+FK3	FK2+FK3	FK1+FK2+FK3
Off Left	0.00%	50.00%	0.00%	50.00%	0.00%	50.00%	50.00%	50.00%
Play	100.00%	0.00%	40.00%	70.00%	50.00%	70.00%	70.00%	70.00%
Pointillism	0.00%	33.33%	66.67%	55.56%	33.33%	66.67%	60.00%	60.00%
Hit	0.00%	33.33%	0.00%	42.86%	0.00%	0.00%	0.00%	0.00%
More Space Fader	0.00%	0.00%	100.00%	50.00%	50.00%	50.00%	50.00%	50.00%
Scanning Right	33.33%	40.00%	40.00%	46.16%	42.86%	45.45%	46.16%	45.45%
Scanning Left	0.00%	0.00%	75.00%	46.16%	45.45%	100.00%	100.00%	75.00%
Whole Group	0.00%	0.00%	75.00%	100.00%	80.00%	46.67%	46.67%	46.67%
Rest of the Group	0.00%	0.00%	100.00%	80.00%	57.15%	100.00%	100.00%	100.00%
Match Right	33.33%	40.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
Match Left	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
Minimalism	0.00%	0.00%	0.00%	42.86%	53.34%	42.86%	80.0%	66.67%
Wait	46.67%	0.00%	100.00%	80.00%	100.00%	55.56%	85.72%	88.89%
Long Tone	46.67%	0.00%	0.00%	46.67%	33.33%	45.45%	40.00%	42.86%
Finish your Idea	0.00%	33.33%	44.45%	40.00%	40.00%	44.45%	40.00%	50.00%
Silence	63.64%	0.00%	50.00%	46.67%	50.00%	46.16%	45.45%	45.45%
Off Right	0.00%	0.00%	100.00%	40.00%	100.00%	33.33%	33.33%	33.33%
Précision moyenne	19.04%	13.53%	46.54%	49.23%	43.26%	46.92%	**49.84**%	48.49%

Table 3: Précision obtenue par les caractéristiques extraites à partir du capteur Myo

Geste	Précision			
	Brutes	FM1	FM2	FM1 + FM2
Off Left	33.33%	0.00%	0.00%	0.00%
Play	0.00%	40.00%	33.33%	33.33%
Pointillism	33.33%	55.56%	55.56%	55.56%
Hit	0.00%	42.86%	42.86%	42.86%
More Space Fader	0.00%	0.00%	0.00%	0.00%
Scanning Right	50.00%	50.00%	50.00%	50.00%
Scanning Left	100.00%	100.00%	100.00%	100.00%
Whole Group	0.00%	0.00%	0.00%	0.00%
Rest of the Group	0.00%	0.00%	0.00%	0.00%
Match Right	0.00%	0.00%	0.00%	0.00%
Match Left	33.33%	50.00%	50.00%	50.00%
Minimalism	0.00%	0.00%	0.00%	0.00%
Wait	0.00%	0.00%	0.00%	0.00%
Long Tone	42.86%	33.33%	40.00%	40.00%
Finish your Idea	0.00%	0.00%	0.00%	0.00%
Silence	0.00%	0.00%	0.00%	0.00%
Off Right	0.00%	0.00%	0.00%	0.00%
Précision moyenne	17.23%	**21.87**%	**21.87%**	**21.87**%

Table 4: Précision obtenue par les caractéristiques extraites à partir de la combinaison de deux capteurs (Kinect et Myo)

Geste	Précision			
	K_{Brutes}+ M_{Brutes}	FK2+ FM1+ FM2	FK3+ FM1+ FM2	FK1+FK2+ FK3+FM1+ FM2
Off Left	0.00%	33.33%	0.00%	0.00%
Play	100.00%	0.0%	70.00%	70.00%
Pointillism	0.00%	33.33%	42.86%	40.00%
Hit	0.00%	42.86%	50.00%	0.00%
More Space Fader	0.00%	100.00%	50.00%	50.00%
Scanning Right	0.00%	50.00%	45.45%	45.45%
Scanning Left	0.00%	46.15%	46.15%	46.15%
Whole Group	0.00%	0.00%	50.00%	50.00%
Rest of the Group	0.00%	44.44%	42.86%	44.44%
Match Right	0.00%	0.00%	0.00%	0.00%
Match Left	0.00%	0.00%	0.00%	0.00%
Minimalism	0.00%	0.00%	40.00%	44.00%
Wait	0.00%	0.00%	60.00%	57.14%
Long Tone	0.00%	0.00%	42.86%	0.00%
Finish your Idea	0.00%	0.00%	0.00%	42.86%
Silence	0.00%	33.33%	45.45%	46.15%
Off Right	50.00%	100.00%	50.00%	50.00%
Précision moyenne	8.82%	28.44%	**36.97%**	34.51%

Learning). Le temps de calcul sera réduit en utilisant des algorithmes plus rapides et plus performants de type K-means[28]. Nous continuerons également à étendre la bibliothèque de gestes pour proposer une plus grande expressivité au SoundPainter et à la communauté des chercheurs. Finalement, nous évaluerons notre système avec des Soundpainters pour leur donner la possibilité de réaliser des compositions artistiques improvisées dans des performances en live.

Finalement, nous pensons que ces premiers résultats sont intéressants sur le plan plus général de la reconnaissance de geste effectués avec tout le corps.

Table 5: Évaluation des précision obtenues avec les cadences d'acquisition de chaque capteur

Kinect® (FK3) (30 Hz)	Myo™ (FM1 + FM2) (200 Hz)	PR_{avg}
30 Hz	30 Hz	**36.97**%
200 Hz	200 Hz	33.62%

REMERCIEMENTS

Les auteurs remercient Giorgio SOGGIU, Rafi SANTOURIAN et Sévan RETIF pour leur collaboration à la création de la base de données de gestes Soundpainting.

Table 6: Temps de calcul moyenne (ms) de notre système pour chaque caractéristique extraite

Device	Charac.	Extrac. (ms)	Réduc. (ms)	Predic. (ms)	Total (ms)
Kinect®	FK1+FK2+FK3	5 ± 1	301 ± 1	21 ± 1	336 ± 1
Myo®	FM1+FM2	2 ± 1	301 ± 1	10 ± 1	324 ± 1
Kinect® + Myo®	FK1+FK2+FK2 +FM1+FM2	8 ± 1	301 ± 1	23 ± 1	349 ± 1

RÉFÉRENCES

[1] 2001. *Interpolation*. John Wiley and Sons, Ltd, Chapter 10, 297–332. https://doi.org/10.1002/0470841621.ch10 arXiv:https://onlinelibrary.wiley.com/doi/pdf/10.1002/0470841621.ch10

[2] 2019. Leap Motion. https://www.leapmotion.com/.

[3] Levent Aksoy, Eduardo Costa, Paulo Flores, and José Monteiro. 2012. Design of Low-complexity Digital Finite Impulse Response Filters on FPGAs. In *Proceedings of the Conference on Design, Automation and Test in Europe (DATE '12)*. EDA Consortium, San Jose, CA, USA, 1197–1202. http://dl.acm.org/citation.cfm?id=2492708.2493004

[4] Z. Arief, I. A. Sulistijono, and R. A. Ardiansyah. 2015. Comparison of five time series EMG features extractions using Myo Armband. In *2015 International Electronics Symposium (IES)*. 11–14. https://doi.org/10.1109/ELECSYM.2015.7380805

[5] Maryam Asadi-Aghbolaghi, Albert Clapés, Marco Bellantonio, Hugo J. Escalante, Victor Ponce-López, Xavier Baró, Isabelle Guyon, Shohreh Kasaei, and Sergio Escalera. 2017. A Survey on Deep Learning Based Approaches for Action and Gesture Recognition in Image Sequences *(FG'17)*. 476–483. https://doi.org/10.1109/FG.2017.150

[6] Ming-Shaung Chang, Jung-Hua Chou, and Chun-Mu Wu. 2009. Establishing A Natural HRI System for Mobile Robot through Human Hand Gestures *(9th IFAC)*. 9–12.

[7] Chen Chen, Kui Liu, and Nasser Kehtarnavaz. 2016. Real-time Human Action Recognition Based on Depth Motion Maps. *J. Real-Time Image Process.* 12, 1 (2016), 155–163. https://doi.org/10.1007/s11554-013-0370-1

[8] Enea Cippitelli, Samuele Gasparrini, Ennio Gambi, and Susanna Spinsante. 2016. A Human Activity Recognition System Using Skeleton Data from RGBD Sensors. *Computational Intelligence and Neuroscience* 2016 (2016), 14. https://doi.org/10.1155/2016/4351435

[9] Alexis Clay, Gaël Domenger, Julien Conan, Axel Domenger, and Nadine Couture. 2014. Integrating Augmented Reality to Enhance Expression, Interaction & Collaboration in Live Performances: a Ballet Dance Case Study *(ISMAR'14)*. IEEE, 21–29.

[10] David Dalmazzo and Rafael Ramirez. 2017. Air Violin: A Machine Learning Approach to Fingering Gesture Recognition. In *Proceedings of the 1st ACM SIGCHI International Workshop on Multimodal Interaction for Education (MIE 2017)*. ACM, New York, NY, USA, 63–66. https://doi.org/10.1145/3139513.3139526

[11] David Dalmazzo, Simone Tassani, and Rafael Ramírez. 2018. A Machine Learning Approach to Violin Bow Technique Classification: A Comparison Between IMU and MOCAP Systems. In *Proceedings of the 5th International Workshop on Sensor-based Activity Recognition and Interaction (iWOAR '18)*. ACM, New York, NY, USA, Article 12, 8 pages. https://doi.org/10.1145/3266157.3266216

[12] Shabnam Sadeghi Esfahlani, Bogdan Muresan, Alireza Sanaei, and George Wilson. 2018. Validity of the Kinect and Myo armband in a serious game for assessing upper limb movement. *Entertainment Computing* 27 (2018), 150–156.

[13] Yona Falinie A. Gaus, Temitayo Olugbade, Asim Jan, Rui Qin, Jingxin Liu, Fan Zhang, Hongying Meng, and Nadia Bianchi-Berthouze. 2015. Social Touch Gesture Recognition Using Random Forest and Boosting on Distinct Feature Sets. In *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction (ICMI '15)*. ACM, New York, NY, USA, 399–406. https://doi.org/10.1145/2818346.2830599

[14] David Gomez, Irvin Dongo, and Nadine Couture. 2019. Automatic Recognition of Soundpainting for the Generation of Electronic Music Sounds *(NIME'19)*. IEEE, 21–29.

[15] Patrice Guyot and Thomas Pellegrini. 2016. Vers la transcription automatique de gestes du soundpainting pour l'analyse de performances interactives *(JIM'16)*. Albi, France, pp. 118–123. https://hal.archives-ouvertes.fr/hal-01530246

[16] Zhenxiang Huang, Bo Peng, and Juan Wu. 2012. Research and Application of Human-Computer Interaction System Based on Gesture Recognition Technology *(CCTA'12)*, Vol. AICT-392. Springer, 210–215.

[17] Feng Jiang, Shengping Zhang, Shen Wu, Yang Gao, and Debin Zhao. 2015. Multi-layered Gesture Recognition with Kinect. *J. Mach. Learn. Res.* 16, 1 (2015), 227–254.

[18] D. Lahat, T. Adali, and C. Jutten. 2015. Multimodal Data Fusion: An Overview of Methods, Challenges, and Prospects. *Proc. IEEE* 103, 9 (Sep. 2015), 1449–1477. https://doi.org/10.1109/JPROC.2015.2460697

[19] W. Li, J. Hou, and L. Yin. 2014. A classifier fusion method based on classifier accuracy. In *2014 International Conference on Mechatronics and Control (ICMC)*. 2119–2122.

[20] Andy Liaw and Matthew C. Wiener. 2007. Classification and regression by Random Forest.

[21] James B. MacQueen. 1967. Some methods for classification and analysis of multivariate observations.

[22] Morgan McCullough, Hong Xu, Joel Michelson, Matthew Jackoski, Wyatt Pease, William Cobb, William Kalescky, Joshua Ladd, and Betsy Williams. 2015. Myo Arm: Swinging to Explore a VE. In *Proceedings of the ACM SIGGRAPH Symposium on Applied Perception (SAP '15)*. ACM, New York, NY, USA, 107–113. https://doi.org/10.1145/2804408.2804416

[23] Leandro Miranda, Thales Vieira, Dimas Martinez, Thomas Lewiner, Antonio W. Vieira, and Mario F. M. Campos. 2012. Real-Time Gesture Recognition from Depth Data through Key Poses Learning and Decision Forests *(SIBGRAPI'12)*. 268–275. https://doi.org/10.1109/SIBGRAPI.2012.44

[24] Tobias Mulling and Mithileysh Sathiyanarayanan. 2015. Characteristics of Hand Gesture Navigation: A Case Study Using a Wearable Device (MYO). In *Proceedings of the 2015 British HCI Conference (British HCI '15)*. ACM, New York, NY, USA, 283–284. https://doi.org/10.1145/2783446.2783612

[25] A. S. Nikam and A. G. Ambekar. 2016. Sign language recognition using image based hand gesture recognition techniques. In *2016 Online International Conference on Green Engineering and Technologies (ICGET)*. 1–5. https://doi.org/10.1109/GET.2016.7916786

[26] Sangjun O., Rammohan Mallipeddi, and Minho Lee. 2015. Real Time Hand Gesture Recognition Using Random Forest and Linear Discriminant Analysis. In *Proceedings of the 3rd International Conference on Human-Agent Interaction (HAI '15)*. ACM, New York, NY, USA, 279–282. https://doi.org/10.1145/2814940.2814997

[27] Iason Oikonomidis, Nikolaos Kyriazis, and Antonis A Argyros. 2011. Efficient model-based 3D tracking of hand articulations using Kinect.. In *BmVC*, Vol. 1. 3.

[28] Peter O. Olukanmi, Fulufhelo Vincent Nelwamondo, and Tshilidzi Marwala. 2018. k-Means-Lite: Real Time Clustering for Large Datasets. *2018 5th International Conference on Soft Computing and Machine Intelligence (ISCMI)* (2018), 54–59.

[29] Thomas Pellegrini, Patrice Guyot, Baptiste Angles, Christophe Mollaret, and Christophe Mangou. 2014. Towards Soundpainting Gesture Recognition. In *Proceedings of the 9th Audio Mostly: A Conference on Interaction With Sound (AM '14)*. ACM, New York, NY, USA, Article 18, 6 pages. https://doi.org/10.1145/2636879.2636899

[30] S. Shahriar, A. Siddiquee, T. Islam, A. Ghosh, R. Chakraborty, A. I. Khan, C. Shahnaz, and S. A. Fattah. 2018. Real-Time American Sign Language Recognition Using Skin Segmentation and Image Category Classification with Convolutional Neural Network and Deep Learning. In *TENCON 2018 - 2018 IEEE Region 10 Conference*. 1168–1171. https://doi.org/10.1109/TENCON.2018.8650524

[31] Toby Sharp, Cem Keskin, Duncan Robertson, Jonathan Taylor, Jamie Shotton, David Kim, Christoph Rhemann, Ido Leichter, Alon Vinnikov, Yichen Wei, Daniel Freedman, Pushmeet Kohli, Eyal Krupka, Andrew Fitzgibbon, and Shahram Izadi. 2015. Accurate, Robust, and Flexible Real-time Hand Tracking. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems (CHI '15)*. ACM, New York, NY, USA, 3633–3642. https://doi.org/10.1145/2702123.2702179

[32] Direk Sueaseenak, Thunchanok Uburi, and Paphawarin Tirasuwannarat. 2017. Optimal Placement of Multi-Channels sEMG Electrod for Finger Movement Classification. In *Proceedings of the 2017 4th International Conference on Biomedical and Bioinformatics Engineering (ICBBE 2017)*. ACM, New York, NY, USA, 78–83. https://doi.org/10.1145/3168776.3168802

[33] Ahmed Taha, Hala H Zayed, ME Khalifa, and El-Sayed M El-Horbaty. 2015. Human activity recognition for surveillance applications *(ICIT'15)*. 577–586.

[34] A. Tanaka, B. Di Donato, and M. Zbyszyński. 2019. Designing Gestures for Continuous Sonic Interaction. In *Proceedings of the New Interfaces for Musical Expression (NIME) (NIME 2019)*. Porto Alegre, Brasil.

[35] Cagri Terdem, Katja Henriksen Schia, and Alexander Refsum Jensenius. 2019. Vrengt: A Shared Body-Machine Instrument for Music-Dance Performance. In *Proceedings of the New Interfaces for Musical Expression (NIME) (NIME 2019)*. Porto Alegre, Brasil.

[36] Walter Thompson. 2006. *Soundpainting: the art of live composition. Workbook 1.*

[37] Walter Thompson. 2009. *Soundpainting: the art of live composition. Workbook 2.*

[38] Walter Thompson. 2014. *Soundpainting: the art of live composition. Workbook 3.*

[39] Robert Y. Wang and Jovan Popović. 2009. Real-time Hand-tracking with a Color Glove. In *ACM SIGGRAPH 2009 Papers (SIGGRAPH '09)*. ACM, New York, NY, USA, Article 63, 8 pages. https://doi.org/10.1145/1576246.1531369

[40] Ruiduo Yang and Sudeep Sarkar. 2006. Gesture Recognition Using Hidden Markov Models from Fragmented Observations *(CVPR '06)*. IEEE Computer Society, 766–773.

[41] Songyang Zhang, Xiaoming Liu, and Jun Xiao. 2017. On Geometric Features for Skeleton-Based Action Recognition Using Multilayer LSTM Networks *(WACV'17)*. 148–157. https://doi.org/10.1109/WACV.2017.24