Détection d'Intention dans un Environnement de Conception en Réalité Virtuelle

R. Guillaume, X. Laville, Y. Baudin
Airbus Operations SAS
316 route de Bayonne
31060 Toulouse – France
romain.r.guillaume@airbus.com

R. Guillaume, J. Pailhès, E. Gruhier
I2M ENSAM Bordeaux
Esplanade d'Arts et Métiers
33400 Talence – France
romain.guillaume@ensam.eu

R. Lou
Arts et Métiers Institute of Technology
LISPEN, HESAM Université, UBFC
71100 Chalon-Sur-Saône, France

Résumé— Ce papier présente une méthode qui devra être capable de traduire les intentions de concepteurs appartenant à différents corps de métier en instructions pour un système immersif basé sur la réalité virtuelle. Ce dernier sera utilisé pour les phases d'allocation d'espace et d'architecture. La méthode est développée dans le cadre de la conception préliminaire d'avions chez l'avionneur Airbus durant les phases d'allocation d'espace et de création d'exigences techniques. Pour répondre à la complexité et la diversité des intentions pouvant être traduites dans le logiciel, nous proposons un modèle en quatre parties (adaptation, extraction d'intentions, extraction de paramètres et traduction en instructions) ainsi que la logique de construction et les différentes difficultés à prendre en compte pour répondre aux besoins de l'outil notamment en termes de performance en temps réel.

Mots-clés— conception, réalité virtuelle, machine learning, intention, langage corporel

I. INTRODUCTION

L'optimisation des processus de conception est au cœur de nombreuses recherches de par la volonté de l'industrie aéronautique de réduire au maximum le nombre et la durée de leurs cycles de production. Ceux-ci souhaitent rester compétitifs sur le marché en proposant des produits en phase avec les derniers progrès technologiques tout en intégrant au plus tôt un maximum de problématiques liées au développement du produit comme les problématiques d'industrialisation ou de maintenance par exemple. Cette constatation est d'autant plus vraie que les produits deviennent de plus en plus complexes, avec des durées de développement qui peuvent parfois dépasser la décennie. La collaboration entre les différents corps de métiers et la simplicité d'utilisation des outils de conception sont des axes de recherches privilégiés qui ont poussé des entreprises comme Airbus à proposer des outils de conception en réalité virtuelle pour les phases d'architecture et d'allocation d'espace de leurs projets [1]. L'utilisation de cette interface immersive sur des cas d'étude a permis de réduire significativement le temps de cycle de conception, faisant par exemple passer la durée de définition d'un harnais électrique de six mois à trois semaines.

L'immersion à échelle 1:1 permet d'avoir une meilleure appréhension des produits, concepts, instructions représentés, en permettant au besoin une superposition entre le réel et cette représentation. Cependant, la mise en place de ces dispositifs nécessite souvent un équipement spécial (casque, lunettes, manettes…) et/ou un apprentissage du langage et de l'ergonomie de l'outil (organisation des menus, commandes boutons et joysticks de la manette…) pour traduire convenablement les intentions de conception des acteurs. Ainsi, l'intermédiation d'un contrôleur est très souvent nécessaire pour traduire les intentions en actions, d'autant plus lorsque celles-ci deviennent diversifiées et complexes. Il existe aujourd'hui différentes manières d'opérer la réalité virtuelle [2] comme les casques de réalité virtuelle ou les salles dotées de larges écrans stéréoscopiques.

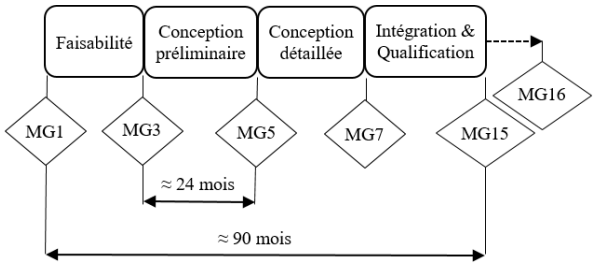

Figure 1. PLANNING DE DEVELOPPEMENT DE PRODUIT AIRBUS

Un programme Airbus est planifié sur 90 mois répartis en 16 "Maturity Gates" (MG) allant de MG1 à MG16 (Figure 1). Notre étude se focalise ainsi sur la période entre MG3 et MG5 correspondant à la phase d'architecture et qui dure environ 24 mois. Cette phase a pour but de générer les différentes allocations d'espace et exigences pour chacun des groupes fonctionnels du produit (électrique, pneumatique…) à partir des objectifs du programme ou "Top Program Objectives" (TPO) livrés à la fin de MG3. La vue du produit passe alors progressivement d'une vue fonctionnelle à une vue physique.

Nous cherchons ainsi à mettre au point un outil de détection d'intentions de conception pour les opérations d'architecture et d'allocation d'espace ne nécessitant pas d'équipement ni d'apprentissage de l'outil par l'utilisateur. Il devra répondre aux besoins des architectes et permettre de renforcer la collaboration des différents acteurs (structure, hydraulique, pneumatique, électrique, système…) dans une même scène de réalité virtuelle. Un exemple de scène serait deux architectes système et structure collaborant des emplacements et dimensions des meubles pour optimiser la répartition des masses de la structure de l'avion. Chacun manipulerait la ou les pièces avec des gestes, en assignant des fonctions à chaque

sous-ensemble notamment grâce à l'interprétation des paroles des architectes.

L'état de l'art autour de la détection d'intentions et des technologies fréquemment rencontrées pour y parvenir est détaillé dans la Section (II), nous proposerons une méthodologie de recherche développée dans la Section (III).

II. ÉTAT DE L'ART

Dans cette section, nous verrons d'abord les techniques couramment employées pour catégoriser de l'information en général (A), puis sur les tentatives empiriques de classification du langage corporel (B) et enfin sur les méthodes utilisées pour détecter l'intention sans intervention humaine (C).

A. Catégorisation

Par définition, la catégorisation est l'action d'organiser des choses par groupes de mêmes caractéristiques [3]. Par exemple, on peut catégoriser arbitrairement les nombres entiers en multiples de 2 et multiples de 3. On notera que suivant nos choix de catégories, certaines données n'appartiennent à aucune catégorie (par exemple le chiffre 5) tandis que d'autres appartiennent à plusieurs d'entre elles (par exemple le chiffre 6). Par ailleurs, l'extraction de certaines caractéristiques peut se réduire à un test de divisibilité, là où d'autres (détection de sourire) nécessitent généralement des outils d'apprentissage pour réussir correctement leur détection. La première catégorie ne nécessitant pas d'outil particulier pour organiser les données, nous nous intéresserons ici plus à la seconde. En effet, la résolution des problèmes avec de l'apprentissage induit des incertitudes sur les résultats qu'il faut pouvoir exploiter. Par exemple, un algorithme ayant appris à détecter si une critique de film est positive en étudiant la distribution des mots dans un message peut se tromper si la tournure est ironique.

Certains algorithmes d'apprentissage ont été pensés pour faire de la classification. Dans le cas des arbres de décisions, les comparaisons faites au niveau des nœuds aiguillent l'objet à classer vers une certaine catégorie, là où les machines à vecteur support segmentent l'espace des caractéristiques en deux (ou plus) catégories avec des frontières (ou marges). D'autres techniques conçues pour s'adapter à une plus grande variété de problème comme les réseaux de neurones ont nécessité la création de technique particulière pour ce type de problème. La plus utilisée d'entre elle étant le "one-hot-encoding" [4]. Cela consiste à obtenir un vecteur de probabilité en sortie de notre réseau afin de choisir la classe la plus probable. Il est également d'usage de prévoir une composante qui n'appartient à aucune classe. On notera également le "ordinal encoding" [4] qui consiste à faire correspondre chaque classe à un entier. Cette méthode induits une relation d'ordre entre les classes rarement désirée. Cependant, ces techniques ne sont pas suffisantes pour traiter les données appartenant à plusieurs classes.

Plusieurs stratégies existent pour résoudre le problème des données à plusieurs classes ou "multi-label" [5]. Les deux grandes familles de stratégie consistent à adapter quand cela est possible un algorithme existant (comme la technique des k voisins les plus proches), ou de transformer le problème "multi-label" en plusieurs problèmes "single-label" du paragraphe précédent. L'étude comparative recommande l'utilisation des algorithmes RF-PCT (Random forest of predictive clustering trees), HOMER (Hierarchy Of Multi-label ClassifiERs), BR (Binary Relevance) ou CC (Classifier Chaining).

B. Études empiriques des liens entre le langage corporel et l'intention

La plupart des études réalisées sur le langage corporel se focalisent sur la posture et la parole [6] [7] [8] [9]. On notera l'existence d'études sur les interfaces cerveau machine [10] sur lesquelles nous ne nous attarderons pas ici, celles-ci traitant souvent plus de la prédiction de mouvement (d'un membre manquant par exemple), que de la prédiction d'intention.

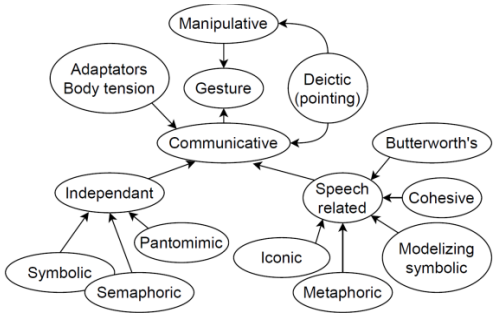

Figure 2. CATEGORISATION DES MOUVEMENTS (VULETIC ET AL., 2019)

Une revue systématique de la littérature sur la classification des mouvements [8] permet de dresser une ontologie des gestes (Catégorisation des mouvements (Vuletic et al., 2019)) réalisés par un humain. On distingue rapidement deux grandes catégories de gestes : les gestes de manipulation et ceux de communication, cette dernière catégorie étant elle-même subdivisée en deux autres catégories : les gestes indépendants de la parole et ceux qui y font directement référence. Hormis le couplage entre la parole et la posture de la personne qu'on retrouve souvent dans la littérature [9], on notera que la majorité des gestes répertoriés sont métaphoriques. En effet, même dans le cadre des gestes de manipulation, la personne est souvent amenée à manipuler un objet virtuel qu'elle imagine ou visualise entre ses mains. D'autre part, certains gestes du quotidien comme un "pouce vers le haut" sont directement liés à un langage gestuel appris par la personne ; certains de ces gestes peuvent être considérés comme universels (symboliques) alors que d'autres sont réservés à un groupe de personne ayant eu un apprentissage spécifique de ces gestes (sémantiques) comme par exemple le langage de communication des plongeurs. Enfin, les gestes dit "adaptateurs" ne contiennent aucun sens et ne servent qu'à relâcher les tensions musculaires.

Une autre étude a été menée sur l'élicitation de gestes et de paroles dans un contexte de conception assistée en 3D [6]. Un expert est alors chargé de réaliser l'action dans un logiciel de conception et d'évaluer à posteriori la simplicité de l'échange avec le participant pour réaliser les tâches. On notera que l'expert n'était pas perceptible par l'utilisateur afin d'éviter les échanges sociaux qui auraient pu nuire à l'étude. Par ailleurs, l'auteur insiste sur la diversité du panel de participants car les différences culturelles et de connaissances techniques pourraient influencer les résultats et mener à des conclusions non généralisables à toute la population [11]. L'expert note à

posteriori la compréhensibilité de l'interaction avec chaque participant. Les résultats montrent une meilleure compréhension avec les gestes et la parole notamment auprès des personnes n'ayant pas ou peu de connaissances en ingénierie mécanique (comme on pourrait avoir une intervention marketing dans un projet de personnalisation de cabine). De plus, les gestes sont catégorisés suivant différents critères puis croisées avec les actions auxquelles ils faisaient référence. Ces critères sont les mouvements relatifs des bras, la posture de la main, et le lexique oral. Les distributions statistiques par actions obtenues pour chacun de ces critères sont significativement différentes entre certaines actions, et en font un bon moyen de différencier l'action désirée par la personne.

C. Automatisation de la détection d'intention

1) Détection de l'intentionnalité

Les techniques les plus simples sont capables de détecter la conscience d'un évènement ou la volonté d'interaction d'une personne avec un objet particulier de la scène [12] [13]. Ces techniques ont la particularité d'être basées sur des heuristiques empiriques très simples et facilement compréhensibles.

Une étude sur la conscience partagée [12] montre que le suivi de la dilatation pupillaire d'une personne se révèle être un bon indicateur pour mesurer la concentration d'une personne sur un objet de la scène. L'expérimentation ayant traquée la dilatation pupillaire des participants s'est déroulée avec les participants assis devant un écran d'ordinateur équipé de caméras RGB et infrarouges.

D'autres études exploitent l'orientation du visage des participants. Une étude sur le partage collaboratif sur grand écran [13] monitore le temps passé à fixer une même zone de l'écran et infère une volonté d'interagir après un certain lapse de temps. Cette méthode présente l'inconvénient d'être fortement dépendante de la complexité de l'information consultée sur l'écran. Par ailleurs, la précision obtenue à partir d'une caméra stéréoscopique placée au-dessus de l'écran est limitée : un écran de 4x3m (LxH) a pu être subdivisé en 6x3 cellules au maximum avec un participant situé à 6m de l'écran. On note ainsi une meilleure précision sur l'orientation horizontale que verticale du visage. L'étude conclut aussi que l'utilisation d'un système de suivi du regard n'était pas adaptée aux écrans larges à cause de leur sensibilité à la lumière et la proximité du visage avec le capteur pour maintenir une précision acceptable. Une autre étude sur les interactions opérateur robot en atelier [14] utilise l'orientation de la tête et la direction du regard comme base de prédiction de la trajectoire d'un opérateur. La méthode décrite nécessite l'utilisation de lunettes de réalité augmentée pour les opérateurs ainsi que l'installation de caméras RGB dans l'atelier.

Ces différentes techniques parviennent à déterminer correctement l'intention des participants dans un contexte où le vocabulaire d'intention est très restreint comme prédire si un sujet a l'intention d'aller à gauche ou à droite sans plus de précisions. Elles deviennent limitées lorsque différentes intentions peuvent être portées sur un même objet ou une même région de l'espace.

2) Discriminants pour le choix de l'intention

Nous avons vu dans la partie expérimentale une étude ayant révélé en outre que la posture des mains était un bon indicateur pour le choix de l'intention du participant [6]. Plusieurs études sur la détermination automatique de la posture de la main ont été réalisées [15] [16] [17] [18]. De manière générale cette reconnaissance s'effectue en trois étapes. La première consiste à isoler la ou les mains du reste de l'image. Pour ce faire, il est possible de segmenter sémantiquement l'image [17], de détecter des points d'intérêt en lien avec la main [15], de reconstruire tout ou partie du squelette du sujet [16], ou encore de séparer l'arrière-plan du premier plan en déterminant une distance seuil puis d'étudier la répartition des points au premier plan pour deviner la position de la main [18]. Cette dernière approche, bien que rapide et transparente, suppose que la personne réalise ses gestes à une distance quasiment prédéfinie de la caméra en se tenant dans une position prédéfinie elle aussi (en l'occurrence face à la caméra). Dans un second temps, l'image de la main est traitée pour simplifier la détection de la posture. La zone d'intérêt peut ainsi être segmentée, normalisée [16] ou redressée dans le cas où celle-ci est fortement inclinée [17]. Enfin, la reconnaissance de la posture se fait généralement directement à partir de l'image traitée, soit en utilisant un réseau de neurone de convolution [16] ou dense [18] soit en utilisant la classification par plus proche voisin [7] [17]. Par ailleurs, une de ces études [17] montre une limite liée à la dynamique de la main. En effet, certains mouvements de main sont très rapides (salutations par exemple), et il n'est pas toujours simple de suivre la main et d'étudier convenablement sa posture.

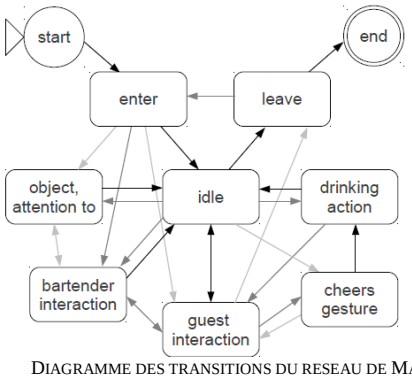

Figure 3. DIAGRAMME DES TRANSITIONS DU RESEAU DE MARKOV CACHE (GASHLER ET AL., 2012)

Une autre étude sur la reconnaissance du comportement en société pour un robot barman [19] propose de se baser sur trois critères pour déterminer l'intention. Premièrement, comme vu précédemment, le modèle va analyser l'orientation du visage du sujet. Puis, un second modèle étudie la posture du sujet. En effet, à l'instar de la pupillométrie vue plus haut, Gashler suggère qu'une personne penchée vers l'objet qu'elle regarde sera plus encline à vouloir interagir avec lui. Plus précisément, la configuration de cette inclinaison vis-à-vis des autres sujets est un indicateur de cette intentionnalité. Enfin, la concaténation des données précédentes est ingérée par un modèle de Markov caché (Diagramme des transitions du réseau de Markov caché (Gashler et al., 2012)) afin de prendre en compte l'historique des dernières intentions du sujet. Comme

pour les solutions précédentes, le vocabulaire des différentes intentions ou des états à repérer est supposé connu.

La détection du squelette d'une ou plusieurs personnes a connu de nombreux progrès notamment en termes d'acuité et de performances [20] [21]. La plupart des méthodes modernes consistent à repérer les articulations puis à chercher à y faire correspondre un modèle cinématique. Ces méthodes ont une bonne résistance à l'occlusion et offrent des performances compatibles avec le temps réel notamment grâce à l'accélération matérielle.

3) Le traitement du langage parlé

Le traitement du langage parlé s'est surtout développé à partir de la fin des années 90 [22] avec l'avènement de l'apprentissage par ordinateur. En effet, les précédentes tentatives de compréhension du langage étaient uniquement basées sur un couplage entre l'analyse grammaticale et lexicale du langage [22] [23]. Cette dernière approche avait l'inconvénient d'être souvent spécifique à une langue ou famille de langue et leurs logiques souvent complexes étaient difficiles à implémenter dans un programme informatique.

Ce secteur de recherche souvent appelé NLP pour "Natural Language Processing" regroupe l'ensemble des traitements de la langue et pas seulement la recherche d'une interprétation précise d'une phrase [24]. On citera par exemple les utilisations suivantes :

- Détection de la langue ;
- Détection du thème ;
- Détection de l'auteur ;
- Etiquetage grammatical d'une partie d'un discours (nature et fonction des mots de la phrase) ;
- Reconnaissance d'entités propres (noms propres, nom de ville, de sociétés…) ;
- Désambiguïsation : certaines tournures de phrase peuvent facilement être mal interprétées comme la phrase type en anglais : "One morning I shot an elephant in my pajamas" ;
- Extraction de l'arbre syntaxique d'une phrase ;
- Traduction automatique.

Ces différentes utilisations du NLP ont été rendues possibles grâce à la réduction de la complexité de programmation d'une part car la grammaire et le lexique de la langue sont (si nécessaires) directement contenus et appris dans les paramètres du modèle, mais aussi grâce au développement de nouvelles capacités de calculs des ordinateurs et plus précisément l'accélération matérielle. D'autre part, les données d'entrée de la plupart des algorithmes de NLP actuels sont sous forme de texte, les entrées sonores étant généralement converties en texte avant leur exploitation.

Des algorithmes de NLP plus complexes combinent plusieurs de ces disciplines ainsi que des mécanismes de "deep-learning" pour extraire des caractéristiques plus complexes comme l'intention de l'utilisateur dans le cas d'une interaction avec un assistant vocal par exemple [25] [26]. Certaines plateformes mettent à disposition des solutions intégrant plusieurs outils de compréhension du langage [27]. Ces outils permettent d'extraire une intention parmi un vocabulaire choisi, de prendre en compte l'historique des derniers messages en décrivant des scénarii possibles et enfin d'extraire les entités liées à cette intention. Il suffit ensuite de faire correspondre les actions correspondantes à l'intention et aux paramètres détectés.

D. Conclusion

Les méthodes de traitement des entrées vidéo et son décrites dans cette partie nous permettront d'extraire les caractéristiques pertinentes pour notre détection d'intention. Les méthodes de catégorisation nous permettront de même de traiter certains problèmes inhérents à notre étude comme la détection d'une ou plusieurs intentions. Cependant, elles restent encore incapables de prédire une ou plusieurs intentions dans un vocabulaire complexe tout en conservant une précision acceptable sans l'utilisation de contrôleurs supplémentaires.

III. METHODOLOGIE

Cette section détaille les étapes de conception de l'outil. Nous nous attarderons d'abord sur la définition de ses entrées et sortie, puis nous détaillerons dans l'ordre les différentes étapes pour concevoir les quatre composants de l'outil (adaptation, extraction d'intentions, extraction de paramètres et traduction en instruction).

A. Boîte noire

Tout d'abord nous allons chercher à représenter l'outil dans son environnement sous la forme d'une boîte noire (Le modèle boîte noire avec les entrées et sorties identifiées), avec des flux entrants et sortants correspondant aux informations utilisées et délivrées par l'outil. Nous commençons par les sorties qui découlent directement des attendus de l'outil. Dans notre cas, il s'agit d'une liste d'opérations paramétrables (par exemple "Déplacer l'objet X de Y mètres dans la direction Z". La construction du vocabulaire qui sera employé par l'outil est effectuée grâce à différentes interviews avec les opérateurs de la solution actuelle de conception en réalité virtuelle mais aussi avec les différents corps d'architecte qui seront les principaux utilisateurs de l'outil.

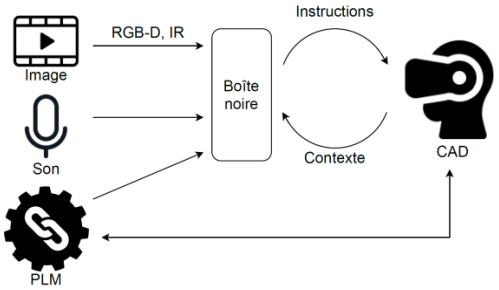

Figure 4. LE MODELE BOITE NOIRE AVEC LES ENTREES ET SORTIES IDENTIFIEES

L'analyse sémantique du vocabulaire ainsi obtenu permettra alors de mettre en évidence les différentes entrées supplémentaires à prévoir en plus des entrées vidéo et son pour notre modèle. Dans l'exemple précédent, le déplacement de "l'objet X" suppose d'avoir une connaissance de cet objet. L'outil devra alors recevoir en entrée la position et certaines métadonnées des différents objets dans l'environnement direct de l'utilisateur.

Les caméras envisagées pour la capture vidéo sont une GoPro Hero 5 qui offre un champ large et une résolution 4K et/ou les caméras de profondeur Intel Realsense d415i et d435i permettant d'obtenir une carte de profondeur jusqu'à une douzaine de mètres ainsi que de capturer des images infrarouges, et sont disponibles au laboratoire Airbus IdLab.

B. *Structure interne*

Dans cette partie, nous détaillerons les différentes étapes proposées pour le développement de notre outil (Figure 5). Nous analyserons alors pour chacune leurs rôles, leurs problématiques respectives et les différentes idées ou concepts pressentis pour y répondre. L'ordre des étapes correspond à l'ordre d'élaboration des parties.

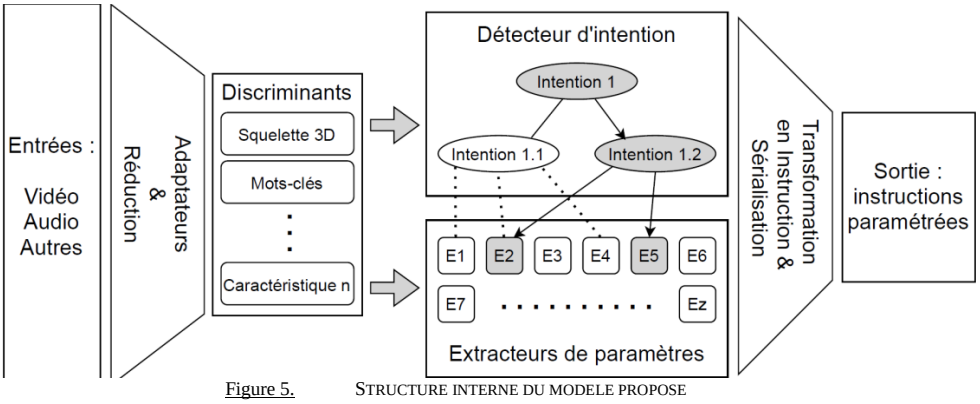

Figure 5. STRUCTURE INTERNE DU MODELE PROPOSE

1) Étape 1 : Vocabulaire d'intention

Cet élément de la solution aura pour objectif de déterminer la ou les intentions de l'utilisateur à partir d'un tenseur de caractéristiques que nous détaillerons dans l'étape 3. Nous proposons tout d'abord de nous concentrer sur les sorties de cet élément, les entrées seront alors déduites de l'étape 3. A l'instar de l'étude des sorties du paragraphe précédent, nous possédons à ce stade de l'étude un vocabulaire d'intention connu. Cette étape définit ainsi la liste des différentes intentions et les paramètres qui y sont associés.

Voici les difficultés identifiées pour cette étape :

- La hiérarchisation et la granularité des intentions : "Dessiner une forme" et "Dessiner un cube" ;
- La combinaison des intentions : on peut assigner une fonction à un objet qu'on est en train de dessiner ;
- La stabilité temporelle de l'intention : une intention choisie doit pouvoir être corrigée sans changer constamment ;
- Le temps réel : la fréquence de calcul doit être limitée.

2) Étape 2 : extraction des paramètres

À partir de la liste des paramètres obtenue à l'étape précédente, des modules seront chargés d'extraire les différents paramètres nécessaires pour chaque intention. Compte tenu de la redondance de paramètres qu'il peut y avoir entre plusieurs intentions comme l'extraction d'une distance par exemple, cette étape est considérée comme étant séparée de la détection d'intention.

Les différentes difficultés répertoriées sont listées ci-après :

- Gérer la diversité de réalisation d'une même intention : on peut par exemple manipuler un objet à une ou deux mains. Dans le premier cas l'orientation sera reliée à celui du poignet de la main alors que dans le second cela sera plus la direction entre les deux paumes de main ;
- Certains paramètres sont en temps réel : à minima la fréquence de calcul de ces derniers devra être limitée ;
- Limiter le nombre d'extracteurs : il faut éviter le recours systématique à la création d'extracteurs de paramètres spécifiques.

3) Étape 3 : adaptateurs et réduction de dimensions

À partir des contraintes des extracteurs de l'étape précédentes, et en s'appuyant sur la littérature pour la détection d'intentions, il devrait être possible à ce stade de lister l'ensemble des signaux faibles et/ou caractéristiques nécessaires pour réaliser les deux premières étapes. Par exemple, dans le cas de l'intention de me déplacer dans la scène, nous aurons besoin de connaître la vitesse et l'orientation de ce déplacement. Dans le cas d'un avatar virtuel, l'intention (implicite) de bouger l'avatar nécessite de connaître les positions et orientations des articulations de l'avatar (à partir de celles de l'utilisateur).

Voici les spécificités techniques trouvées pour cette étape :

- Actualisation des données : des données d'entrée pourront être mises à jour en temps réel comme dans notre exemple tandis que d'autres ne le seront que sous certaines conditions (une position enregistrée par un utilisateur) ou jamais (règle de conception d'un câblage électrique) ;
- Différents modes de fonctionnement de l'outil : la donnée produite doit être agnostique du mode de fonctionnement de l'outil (casque de réalité virtuelle ou PowerWall) ;
- Synchronisation par utilisateur : pour une utilisation sur PowerWall où plusieurs personnes sont amenées à évoluer sous le champ des caméras et des micros, les entrées vidéo et son devront concorder pour les différents utilisateurs.

4) Étape 4 : traduction en suite d'opérations

Enfin, il faudra traduire l'ensemble des intentions paramétrées que nous aurons identifié en instructions compréhensibles par le logiciel de conception en réalité virtuelle. Cette opération peut être vue comme une opération de sérialisation/analyse. À ce titre, à minima une partie de cette brique fonctionnelle sera spécifique au logiciel, c'est pourquoi

séparer cette brique en deux semble être la meilleure approche en termes d'adaptabilité.

C. *Entraînement*

Compte tenu de la plasticité et la complexité de la tâche, une partie des modules reposera vraisemblablement sur de l'apprentissage par ordinateur. Le découpage en organes de fonctionnement spécifique laisse cependant espérer des modules de complexité localement limitée permettant ainsi un entraînement plus rapide à partir de données d'entraînement restreintes. Par ailleurs, la transparence du modèle permise par cette répartition devrait être suffisante pour entraîner la majeure partie des modules individuellement au travers d'exercices spécifiques. Par exemple, le module d'extraction de dimensions pourra être entraîné séparément de celui d'extraction du squelette de l'utilisateur par exemple.

IV. Conclusion

Ce papier présente la méthode de détection d'intention proposée pour la conception préliminaire en réalité virtuelle qui sera développée dans les prochains mois. Cette méthode est innovante car elle permet d'utiliser une plus grande diversité d'entrées et ainsi de déterminer un panel d'intention beaucoup plus large que ce qu'il est théoriquement faisable avec les entrées vidéo ou son uniquement. Par ailleurs, les acteurs du projet pourront collaborer sans apprentissage des commandes de l'outil en utilisant uniquement leurs voix et gestes. Dans le cadre de l'industrie 4.0, on pourrait imaginer appliquer cette méthodologie à l'analyse de performance ou l'aide à la production en atelier, en prédisant les intentions des différents opérateurs sur place pour leur offrir une expérience de production améliorée et optimisée.

V. Bibliographie

[1] J. BERGOUNHOUX, "La réalité virtuelle se déploie chez Safran et Daher", L'Usine Nouvelle, avr. 2020.

[2] C. TREILLES, "SkyReal équipe les industriels d'une solution VR Cave", ZDNet France, sept. 21, 2020.

[3] C. MCINTOCH, "Cambridge Advanced Learner's Dictionary", 4th éd. Cambridge University Press, 2013.

[4] J. BROWNLEE, "Ordinal and One-Hot Encodings for Categorical Data", Machine Learning Mastery, 2020.

[5] G. MADJAROV, D. KOCEV, D. GJORGJEVIKJ, ET S. DZEROSKI, "An extensive experimental comparison of methods for multi-label learning", Pattern Recognition, vol. 45, no 9, p. 3084 3104, sept. 2012.

[6] S. KHAN ET B. TUNÇER, "Gesture and speech elicitation for 3D CAD modeling in conceptual design", Automation in Construction, vol. 106, p. 102847, 2019.

[7] G. KUTLIROFF ET A. BLEIWEISS, "Method and System for Gesture Classification", US 7,970,176 B2, juin 28, 2011.

[8] T. VULETIC, A. DUFFY, L. HAY, C. MCTEAGUE, G. CAMPBELL, ET M. GREALY, "Systematic literature review of hand gestures used in human computer interaction interfaces", International Journal of Human-Computer Studies, vol. 129, p. 74 94, sept. 2019.

[9] J. EISENSTEIN ET R. DAVIS, "Visual and Linguistic Information in Gesture Classification", ACM international conference on multimodal interfaces, p. 8, oct. 2004.

[10] B. BIOULAC, B. JARRY, ET R. ARDAILLOU, "Rapport 20-06 – Interfaces cerveau-machine : essais d'applications médicales, technologie et questions éthiques", Bulletin de l'Académie Nationale de Médecine, p. 12, nov. 2020.

[11] H. WU ET AL., "Influence of cultural factors on freehand gesture design", International Journal of Human-Computer Studies, vol. 143, p. 102502, nov. 2020.

[12] B. PREBOT, "Représentation partagée et travail collaboratif en contexte C2 : monitoring d'opérateurs en situation simulée de command and control. ", Université de Bordeaux, France, 2020.

[13] D. MARION, "Sharing big display: développement des technologies et métaphores d'interactions nouvelles pour le partage collaboratif d'affichage en groupe ouvert", Université de Bordeaux, France, 2018.

[14] R. T. CHADALAVADA, H. ANDREASSON, M. SCHINDLER, R. PALM, ET A. J. LILIENTHAL, "Bi-directional navigation intent communication using spatial augmented reality and eye-tracking glasses for improved safety in human–robot interaction", Robotics and Computer-Integrated Manufacturing, vol. 61, p. 101830, févr. 2020.

[15] W. ZHAO, Y. BAO, ET H. QU, "Hand gesture understanding by weakly-supervised fusing shallow/deep image attributes", Signal Processing: Image Communication, vol. 82, p. 115760, mars 2020.

[16] J. WANG, T. LIU, ET X. WANG, "Human hand gesture recognition with convolutional neural networks for K-12 double-teachers instruction mode classroom", Infrared Physics & Technology, vol. 111, p. 103464, déc. 2020.

[17] H. LAHAMY, "Real-Time Hand Posture Recognition Using a Range Camera", University Of Calgary, Calgary, 2013.

[18] S. MARCEL, O. BERNIER, ET D. COLLOBERT, "Reconnaisance de la Main pour les Interfaces Gestuelles", COmpression et REprésentation des Signaux Audiovisuels, juin 1999.

[19] A. GASCHLER, S. JENTZSCH, M. GIULIANI, K. HUTH, J. DE RUITER, ET A. KNOLL, "Social behavior recognition using body posture and head pose for human-robot interaction", in 2012 IEEE/RSJ International Conference on Intelligent Robots and Systems, Vilamoura-Algarve, Portugal, oct. 2012.

[20] Z. CAO, G. HIDALGO, T. SIMON, S.-E. WEI, ET Y. SHEIKH, "OpenPose: Realtime Multi-Person 2D Pose Estimation using Part Affinity Fields", IEEE Transactions on Pattern Analysis and Machine Intelligence, mai 2019.

[21] G. PAPANDREOU, T. ZHU, L.-C. CHEN, S. GIDARIS, J. TOMPSON, ET K. MURPHY, "PersonLab: Person Pose Estimation and Instance Segmentation with a Bottom-Up, Part-Based, Geometric Embedding Model", European Conference on Computer Vision, mars 2018.

[22] G. G. CHOWDHURY, "Natural Language Processing", The Annual Review of Information Science and Technology, p. 39, 2003.
[23] R. E. LONGACRE, "The Grammar of Discourse", 1983e éd. Plenum Press New York, 1996.
[24] Y. GOLDBERG, "Neural Network Methods for Natural Language Processing", Synthesis Lectures on Human Language Technologies, vol. 10, no 1, p. 1 309, avr. 2017.
[25] Y. WANG ET AL., "Stacking-Based Ensemble Learning of Self-Media Data for Marketing Intention Detection", Future Internet, vol. 11, no 7, p. 155, juill. 2019.
[26] M. TIWARI, M. MATHIHALLI, K. RANGADURAI, Q. VOHRA, S. DARURU, ET R. N. RAJ, "Natural Language Processing systems and methods", US 2019/0103111 A1, avr. 04, 2019.
[27] T. BOCKLISCH, J. FAULKNER, N. PAWLOWSKI, ET A. NICHOL, "Rasa: Open Source Language Understanding and Dialogue Management", Conference on Neural Information Processing Systems, déc. 2017.