De la trace au modèle : approches de modélisation de l'activité en réalité étendue (XR)

Raihana ALLANI, LUCID, Université de Liège, Belgique, r.allani@uliege.be

Pierre LECLERCQ, LUCID, Université de Liège, Belgique, pierre.leclercq@uliege.be

Résumé

Les dispositifs de réalité étendue (XR) transforment les conditions dans lesquelles l'activité humaine se déploie. Ils articulent espaces physiques et numériques en favorisant l'interaction avec des objets et des acteurs distribués dans les deux environnements et en engagent l'utilisateur dans une expérience immersive. Modéliser cette activité soulève une question méthodologique : comment transformer les phénomènes observés en modèles pertinents sans réduire la complexité de l'activité en XR ? Notre recherche tire parti d'une spécificité majeure des dispositifs XR. Leur instrumentation native produit en continu des flux multimodaux de traces décrivant l'activité avec une finesse inédite. Cet article propose une typologie analytique des principales approches qui transforment ces traces en modèles, sans viser l'exhaustivité des outils existants. Trois logiques sont distinguées : le codage et l'annotation multimodale, la modélisation par variables mesurables, et la reconnaissance automatique de l'activité. Une discussion transversale en dégage les compromis méthodologiques et la division du travail entre le chercheur et les outils de modélisation. La typologie offre un cadre de positionnement permettant de clarifier les choix de modélisation au regard des spécificités de l'activité en XR et des finalités de la recherche.

Mots-clefs

Modélisation de l'activité, Données multimodales, Réalité étendue, méthode de modélisation, Human Activity Recognition.

1. Introduction

La réalité étendue (XR) regroupe un ensemble de technologies immersives comprenant la réalité virtuelle (VR), la réalité augmentée (AR) et la réalité mixte (MR). Le déploiement croissant de ces technologies dans des domaines aussi divers que l'industrie, la santé, la formation et particulièrement dans les musées, transforme les conditions dans lesquelles l'activité humaine se déploie. Celle-ci ne se déroule plus uniquement dans un espace matériel ni exclusivement dans un espace numérique, mais dans une configuration hybride où coexistent le corps de l'utilisateur, des objets physiques et des objets virtuels. Cette hybridation s'accompagne d'une instrumentation native car les dispositifs XR enregistrent, en flux continu, les interactions et les mouvements de l'utilisateur et peuvent même capturer des signaux physiologiques.

Modéliser l'activité à partir de ces traces sert différents objectifs selon les visées de la recherche, qu'il s'agisse de caractériser des pratiques d'usage, d'évaluer une expérience immersive ou de concevoir des dispositifs adaptatifs. Si la XR facilite la production de traces riches et hétérogènes, elle rend plus complexe leur transformation en modèles. La question n'est plus seulement celle de l'accès aux données mais celle de leur structuration : *Quelles dimensions retenir ? Comment les organiser ? Quel type de modèle construire* ? Les approches mobilisées pour ce travail sont nombreuses et issues de champs aussi variés (ergonomie, interaction humain-machine, machine learning) mais les logiques qui les structurent restent souvent implicites, ce qui rend difficile leur comparaison et leur mobilisation en fonction des objectifs poursuivis.

Cet article propose de clarifier ces logiques à travers une typologie analytique des approches qui construisent des modèles d'activité à partir des traces effectivement produites par son déploiement. Ce positionnement écarte les approches rétrospectives de restitution de l'activité (autoconfrontation, questionnaires) ainsi que les approches prospectives ou prédictives (agents virtuels, simulations génératives) qui saisissent l'activité par d'autres médiations que les traces. Trois approches sont retenues : le codage et l'annotation multimodale, la mesure d'indicateurs comportementaux, et la reconnaissance automatique de l'activité. Elles sont caractérisées dans les sections 3 à 5 puis articulées dans une discussion transversale qui fait apparaître les compromis méthodologiques que leur choix implique et les limites partagées qui ouvrent autant de directions pour les recherches à venir.

2. Cadres conceptuel et analytique

2.1 Concepts fondamentaux de la modélisation de l'activité en XR

La notion d'activité humaine ne possède pas de définition univoque. Elle varie selon les cadres théoriques et les contextes d'analyse. Dans la tradition issue de Vygotski (1978) et Leontiev (1978), l'activité est pensée comme un système structuré orienté vers un objet, organisé en actions et opérations, et médiatisé par des outils. Cette perspective met en évidence l'inscription sociale et instrumentée de l'action. Barbier (2013) en propose une définition plus large. L'activité est l'ensemble des transformations du monde physique, social ou mental dans lesquelles un sujet est engagé et par lesquelles il se transforme lui-même, sans supposer nécessairement une intention consciente ni une élaboration immédiate de sens. L'activité peut ainsi émerger avant d'être stabilisée dans une signification. Suchman (1987), dans la tradition de l'action située, souligne le caractère ajusté et contingent de l'action dans la situation concrète, tandis que les perspectives de la cognition incarnée et de l'énaction (Varela et al. 1991) insistent sur le rôle central du corps et de l'engagement sensorimoteur dans sa constitution. Latour (2005) propose enfin de considérer l'activité comme distribuée au sein d'un réseau d'actants humains et non humains, où artefacts, dispositifs techniques et configurations spatiales participent à la production même de l'action.

Les environnements de réalité étendue ne redéfinissent pas fondamentalement ces perspectives, mais en accentuent certaines dimensions. En particulier, l'espace n'y constitue plus un simple arrière-plan, il devient une composante active qui contraint, oriente et peut être transformée par l'interaction (Slater and Sanchez-Vives, 2016). Par ailleurs, l'instrumentation native des dispositifs XR rend l'activité traçable à travers des flux multimodaux continus (Rakkolainen et al. 2021). Toutefois, ces traces ne constituent pas l'activité elle-même, elles en représentent une inscription technique nécessairement partielle, centrée sur ses dimensions comportementales et interactionnelles. Entre ces traces et l'expérience vécue s'interpose toujours une opération de sélection et de formalisation. Modéliser l'activité revient dès lors à produire, à partir de ces données, une représentation structurée selon un cadre théorique et un objectif d'analyse (Rozumnyi et al. 2024).

2.2 Positionnement et critères d'analyse

Cet article s'inscrit dans une démarche typologique fondée sur l'identification de principes méthodologiques de modélisation. L'analyse repose sur l'examen de dispositifs représentatifs issus de l'ergonomie et de l'interaction humain-machine en XR. Notre intérêt ne porte cependant pas sur les outils en tant que dispositifs techniques mais sur les méthodes sous-jacentes qui les structurent c'est à dire la manière dont chacun transforme les traces de l'activité en un modèle et les principes méthodologiques qui guident cette transformation. Pour ce faire, nous avons mobilisé un ensemble de critères portant sur la transformation des traces en modèles : la nature des traces mobilisées, le mode de structuration des données, les unités retenues pour modéliser l'activité, le type de modèle produit et la place respective de l'interprétation humaine et du traitement automatisé.

À l'issue de cette analyse, nous avons retenu trois approches de modélisation. La première approche est la « Modélisation par codage et l'annotation multimodale ». Elle construit le modèle en appliquant aux

traces une grille d'annotation définie en amont par le chercheur. La deuxième approche est la « Modélisation par variables mesurables » qui transforme les traces en indicateurs calculés sur les flux produits nativement par la session XR. La troisième approche est la « Reconnaissance automatique de l'activité ». Elle consiste à entraîner un programme à associer automatiquement des configurations de traces à des catégories d'activité à partir d'un corpus préalablement étiqueté. Ces approches se différencient par l'emplacement de l'opération principale de structuration : définie en amont par le chercheur pour le codage, émergente du traitement des indicateurs pour la modélisation par variables mesurables ou résultante d'un processus d'apprentissage pour la reconnaissance.

3. Modélisation par codage et annotation multimodale

3.1 Principes de modélisation

Cette première approche construit le modèle d'activité XR à partir d'un schéma de codage défini en amont. Les traces que produit la session XR (enregistrements audiovisuels, logs d'interaction, trajectoires cinématiques, verbalisations synchronisées) sont segmentées et catégorisées selon un cadre théorique explicite. La structuration du modèle prend ainsi la forme d'un ensemble fini de catégories interprétables ou de séquences organisées d'unités codées. Cette démarche s'appuie sur les traditions de l'éthologie humaine (Bakeman & Gottman, 1997), de la linguistique multimodale (McNeill, 1992) et de l'analyse qualitative assistée par ordinateur dont les outils historiques comme ELAN (Wittenburg et al., 2006) ont stabilisé les principes de grille de codage, de tiers temporellement alignés[1] et de validation inter-juges.

Deux modèles ressortent de cette approche portés par des formalismes distincts. Le premier modèle est séquentiel. Il représente l'activité comme une partition multimodale où chaque ligne porte une modalité (geste, parole, regard) ou un acteur. Cet affichage permet de lire de manière verticale les coïncidences entre lignes et, horizontalement, les enchaînements au sein d'une ligne. Le deuxième modèle est plutôt interactionnel. Il transforme la partition de l'activité en structure statistique (probabilités de transition, patterns temporels récurrents, réseaux de co-occurrences) capturant les régularités d'enchaînement à l'échelle d'un corpus d'épisodes plutôt que la dynamique d'un épisode singulier.

3.2 Use cases

Le système ISA-Immersive Study Analyzer (Lammert et al., 2024) illustre le modèle séquentiel dans sa forme la plus aboutie en XR. Ce système enregistre l'ensemble des actions utilisateur, de la parole et du contexte environnemental d'études en VR sociale. Il permet ensuite à plusieurs chercheurs de coder collaborativement ces traces, au sein même de la scène 3D reconstituée, avec une re-spatialisation audio qui permet de suivre les conversations comme si l'on y était. Comme le montre la figure 1, ISA reprend les principes d'annotation d'ELAN mais les transpose dans un espace immersif. Le chercheur y devient un observateur mobile dans la scène et peut faire des requêtes des événements codés au fil des questions qu'il se pose.

Le modèle interactionnel est illustré par Syiem et al. (2026) dans une étude sur la communication entre géologues qui doivent expliquer à distance des concepts spatiaux complexes. Les auteurs comparent en réalité virtuelle deux outils de dessin, un tableau virtuel plat et un dessin tracé directement dans l'espace 3D, auprès de vingt participants formant dix paires qui discutent de structures géologiques. Les vidéos sont codées selon une grille décrivant les actions accompagnant la parole, comme les gestes de désignation, les déplacements partagés dans l'espace ou les références aux dessins existants. Une analyse en réseaux de co-occurrences est ensuite appliquée pour modéliser les associations entre ces catégories. Le modèle qui en résulte prend la forme d'un réseau pondéré où chaque nœud représente une catégorie codée et chaque arête représente la fréquence avec laquelle deux catégories apparaissent ensemble dans

[1] Un tier est une ligne d'annotation dans laquelle on inscrit, de manière synchronisée avec un enregistrement vidéo ou audio, des événements codés appartenant à une seule catégorie ou portés par un seul acteur.

les conversations. Cette représentation illustrée dans la figure 2 rend visibles des patterns d'articulation entre les interactions qu'une lecture séquentielle ne ferait pas apparaître.

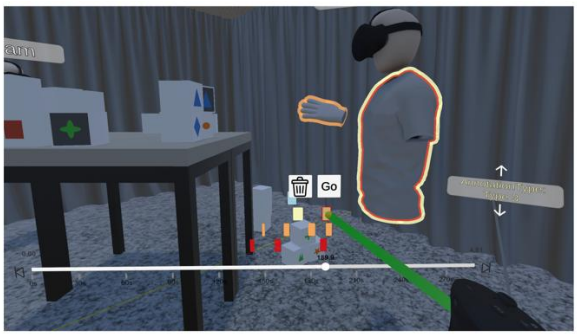


Figure 1. Visualisation et navigation des annotations sur la ligne temporelle du système ISA (Lammert et al., 2024).

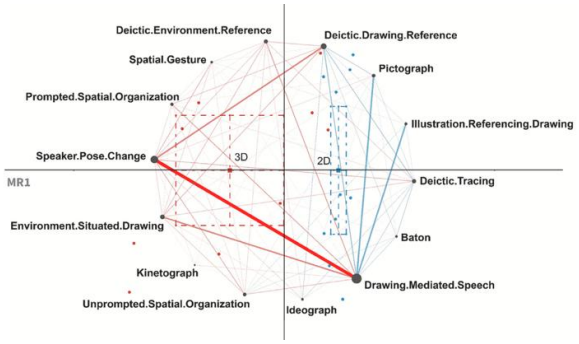


Figure 2. Visualisation de co-occurrence des actions communicatives en 2D et 3D (Syiem et al.,2026).

3.3 Discussion

L'approche par codage et annotation multimodale offre une lecture interprétative de l'activité. Selon son objectif et son cadre théorique, le chercheur attribue des catégories porteuses de sens à des moments précis de la session et articule ces moments en structures. Cette approche se révèle particulièrement précieuse car elle permet l'étude fine de la manière dont les comportements s'articulent entre eux au fil de la session, un niveau d'analyse que les autres approches n'atteignent pas avec la même précision. Sa transposition immersive, illustrée par l'outil ISA, étend cette lecture en permettant un codage collaboratif dans la scène 3D rejouée, ce qui prend tout son sens lorsque la dimension spatiale structure l'activité étudiée.

Une limite structurelle pèse cependant sur cette démarche. La finesse du codage manuel impose une charge humaine que l'exhaustivité des traces immersives rend rapidement prohibitive sur des corpus étendus. L'assistance par IA pour la transcription et le pré-codage apporte une réponse partielle mais déplace la difficulté vers la validation des suggestions de l'IA et la reproductibilité inter-juges.
Un point d'attention mérite également d'être souligné. Le codage produit des unités discrètes dont le chercheur assigne le début et la fin, tandis que les capteurs XR génèrent des flux continus à haute fréquence (position du casque, direction du regard) sans frontières naturelles. Articuler les deux exige de décider à quels moments du flux une catégorie commence et s'achève. Ce travail méthodologique reste faisable mais demande au chercheur d'expliciter les règles de découpage qu'il adopte.

4. Modélisation par variables mesurables

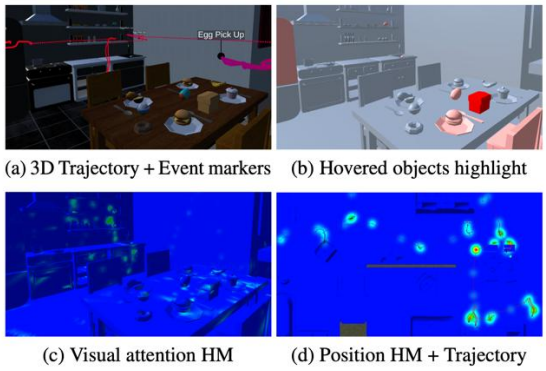
4.1 Principes de modélisation

Cette deuxième approche construit le modèle d'activité XR à partir d'indicateurs calculés sur les flux de traces que la session immersive produit nativement. Plutôt que de catégoriser l'activité en unités interprétatives définies en amont comme dans la première approche, le chercheur identifie les variables qui l'intéressent (distance, durée, fréquence, direction du regard, intensité physiologique) et instrumente l'environnement pour les mesurer en continu. L'élément de base n'est plus l'action identifiée mais l'indicateur calculé. L'interprétation se déplace en aval, dans la lecture des distributions et des dynamiques produites. Si cette approche s'appuie historiquement sur les traditions de la psychophysique expérimentale (Stevens, 1946) et de l'HCI quantitative (Card et al. 1983) qui ont établi qu'une activité peut être caractérisée par un ensemble réduit de variables de performance, elle tire actuellement parti d'une propriété spécifique de la XR. Cette spécificité concerne la production native, par les dispositifs XR, des flux de mesures denses et synchronisés comme le suivi à six degrés de liberté (6DoF), l'eye-tracking intégré et les capteurs physiologiques (EEG) qui sont par défaut le format dans lequel la trace s'enregistre.

Un modèle unique ressort de cette approche. Celui-ci consiste à représenter l'activité par la distribution d'un ou plusieurs indicateurs choisis par le chercheur, dans l'espace, dans le temps ou dans leurs relations mutuelles. L'indicateur représenté peut être une variable directement mesurée comme la position du casque ou la durée d'une fixation du regard, ou une variable latente inférée par un traitement statistique comme la charge cognitive ou l'engagement. Ce qui unifie l'approche n'est pas le format de représentation, qui varie selon les besoins et les indicateurs sélectionnés, mais la démarche elle-même qui consiste à transformer des flux de mesures en représentations, rendant ainsi visibles des dimensions spécifiques de l'activité.

4.2 Use cases

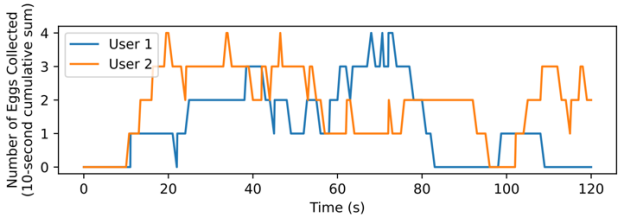
Pour illustrer ce modèle, nous présentons le travail de Javerliat et al. (2024) qui ont développé PLUME, un toolkit open-source d'enregistrement, de rejeu de session et d'analyse de traces comportementales en environnement VR. Ce toolkit capture en continu trois types de traces à savoir les mouvements du corps, le regard et les signaux physiologiques. Il permet le rejeu de la session dans son environnement 3D d'origine et génère des visualisations explorables telles que des cartes de densité, des trajectoires et des courbes physiologiques.
Les auteurs valident cet outil à travers une étude de cas immersive : une tâche de chasse aux œufs de Pâques dans un appartement virtuel, exploré en trois minutes, avec un casque Meta Quest Pro et un capteur cardiaque. À partir des traces, PLUME produit plusieurs représentations notamment une heatmap d'attention visuelle, des trajectoires de déplacement et des signaux physiologiques synchronisés comme illustré dans les figures 3 et 4. Le chercheur peut naviguer entre ces représentations et le rejeu de la session afin d'interpréter les données pour vérifier, par exemple, si une zone de forte attention est liée à un intérêt fort, une difficulté particulière ou une hésitation.



(a) 3D Trajectory + Event markers (b) Hovered objects highlight

(c) Visual attention HM (d) Position HM + Trajectory

Figure 3. Visualisation spatiale des interactions Javerliat et al. (2024).



(a) Collected eggs over time: 2-users comparison.

	Eggs	Grabbed	Most Hovered	Traveled (m)	Teleportation
User 1	23	15	Refrigerator	123.3	67
User 2	15	42	Wine Bottle	88.90	42

(b) Various behavior statistics: 2-users comparison

Figure 4. Comparaison des actions de deux utilisateurs Javerliat et al. (2024).

Quant à la variante inférentielle, nous l'illustrons à travers le travail de Nasri et al. (2024) qui analysent un système de formation VR à un procédé industriel de projection thermique. Les auteurs collectent les données d'eye-tracking de dix-neuf participants et calculent deux indicateurs à partir du regard : la durée des fixations et la dilatation de la pupille (figure 5). Ils entraînent ensuite deux modèles d'apprentissage automatique pour prédire le niveau de charge cognitive ressenti par les participants, mesuré par un questionnaire standardisé d'auto-évaluation. Le modèle produit est une fonction qui, à partir d'un jeu d'indicateurs oculaires, infère un niveau de charge cognitive représentable sous forme de courbe temporelle superposable à la tâche exécutée par l'étudiant.
L'étude montre comment le modèle d'indicateurs peut produire des variables que l'observation directe ne capture pas, ici un état interne de l'acteur, à la condition que le lien théorique entre l'indicateur mesuré et l'état qu'on cherche à mesurer soit solide. Par exemple, la dilatation pupillaire peut réagir aussi bien à la lumière qu'à la charge cognitive.

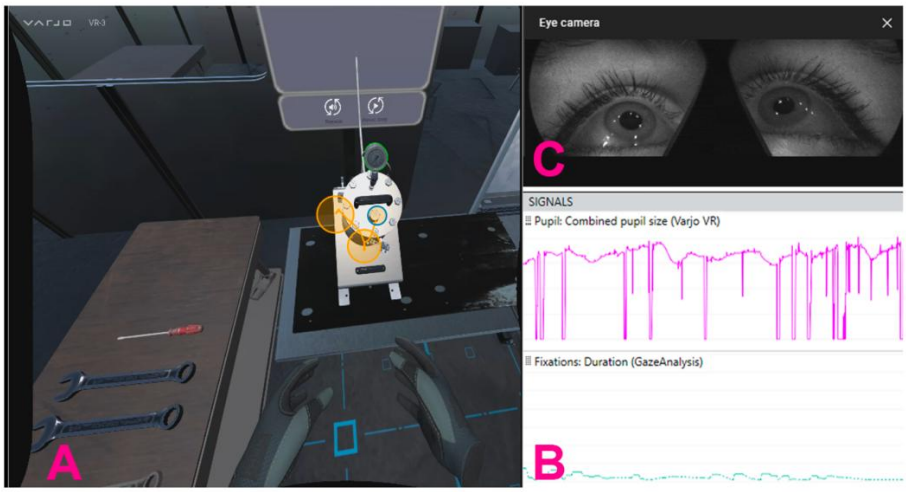


Figure 5. (A) Données de traitement dans iMotions Lab; (B) Signal de dilatation pupillaire et durée des fixations ; (C) Caméra oculaire (Nasri et al., 2024)

4.3 Discussion

Cette approche offre une lecture quantitative de l'activité qui exploite les flux de données produits nativement par les environnements XR. Elle permet de comparer rigoureusement des conditions expérimentales, d'identifier des patterns de distribution spatiale ou temporelle et d'inférer des états internes à partir de signaux mesurables. Les visualisations qu'elle produit à partir des sessions rejouées (cartes de densité projetées en 3D, courbes synchronisées) sont par ailleurs particulièrement lisibles ce qui facilite la communication des résultats vers des audiences non spécialistes.
Plusieurs points d'attention encadrent toutefois sa mise en œuvre. Le choix des variables conditionne d'emblée ce que le modèle rendra visible et ce qu'il masquera. Néanmoins, la densité des traces XR rend tentante une mesure tous azimuts dont la multiplication peut produire des corrélations fortuites. Cette sélection demande un cadrage théorique préalable qui n'est pas toujours formulé explicitement. Par ailleurs, quand l'approche s'oriente vers l'inférence d'états internes, elle dépend de la solidité du lien théorique entre l'indicateur mesuré et l'état qu'on cherche à mesurer. Or ce lien est rarement univoque comme expliqué dans l'exemple de la dilatation pupillaire qui peut réagir aussi bien à la lumière qu'à la charge cognitive. Le modèle inférentiel n'est donc valable que dans la mesure où les conditions de validité de ses indicateurs sont contrôlées. La modalité XR ajoute enfin une vigilance spécifique en AR et MR où les indicateurs spatiaux articulent des référentiels virtuels et physiques incertains. Les modèles qui ignorent cette hétérogénéité risquent de produire des valeurs dont la précision numérique masque l'incertitude réelle.

5. Reconnaissance automatique de l'activité

5.1 Principes de modélisation

Cette troisième approche construit le modèle d'activité XR en déléguant la reconnaissance des épisodes d'activité à un programme entraîné. La place du chercheur s'y déplace car il n'intervient plus pendant l'analyse. Il intervient désormais en amont en constituant un corpus annoté à partir duquel un programme apprend à associer des configurations de traces à des catégories d'activité. Une fois entraîné, le programme peut traiter, sans nouvelle intervention humaine, des sessions qu'il n'a jamais vues. La démarche se rattache à la tradition du Human Activity Recognition (Arshad et al. 2022) stabilisée depuis les années 2000 dans le champ de l'informatique ubiquitaire et transposée plus récemment au contexte immersif.

Cette approche s'appuie sur les deux précédentes plutôt qu'elle ne les remplace. Son corpus d'apprentissage repose sur un travail de codage manuel relevant directement de l'approche présentée en section 3. Ses entrées (*inputs*) sont des caractéristiques calculées sur les traces comme la durée d'une

fixation du regard ou la vitesse d'un déplacement qui sont les mêmes indicateurs qu'analyse l'approche présentée en section 4. Le programme apprend à associer correctement les caractéristiques aux étiquettes du corpus. Cette opération peut être réalisée par différentes architectures allant des classifieurs statistiques classiques aux réseaux neuronaux profonds.

Le programme entraîné produit un modèle de classification d'activité qui, appliqué à une session, génère une suite d'étiquettes le long de son déroulé temporel et découpe ainsi l'activité en épisodes catégorisés. Ce modèle peut être mobilisé selon deux finalités : (1) En différé, sur un enregistrement, il segmente automatiquement les sessions d'un corpus pour permettre la caractérisation de l'activité, la comparaison entre populations ou conditions, ou la détection ciblée d'épisodes d'intérêt sur des volumes que le codage manuel n'atteindrait pas. (2) En ligne, pendant le déroulement de la session, il alimente une boucle d'adaptation qui ajuste l'environnement immersif à ce que l'utilisateur est en train de faire comme l'affichage d'une aide contextuelle, la modification d'un scénario ou le déclenchement d'une consigne. Cette deuxième finalité distingue cette approche des deux précédentes. En effet, elle est la seule à pouvoir agir sur l'environnement au moment même où l'activité se déroule.

5.2 Use cases

Le modèle de la Reconnaissance automatique de l'activité est illustré par Bektaş et al. (2024) via le dispositif GEAR (*Gaze-Enabled Activity Recognition*) dans le cadre d'une étude menée avec vingt participants équipés d'un casque HoloLens 2. Chaque participant réalise successivement trois tâches dans un environnement de bureau : lire un texte affiché en réalité augmentée, inspecter un appareil électronique posé devant lui, et chercher un objet dans la pièce. Pendant ces tâches, le casque enregistre les données de regard, à partir desquelles les auteurs entraînent un programme à reconnaître automatiquement laquelle des trois activités est en cours. Le modèle illustré en figure 6 reconnaît correctement l'activité dans 89,6 % des cas et fonctionne directement sur le casque, ce qui permet d'adapter immédiatement l'assistance affichée : mise en évidence du texte pertinent quand l'utilisateur lit, affichage des propriétés de l'appareil quand il l'inspecte, suggestion d'éléments proches quand il cherche. L'étude met en évidence trois propriétés de l'approche. D'abord sa dépendance à un étiquetage préalable des activités, ensuite la nécessité d'un modèle suffisamment léger pour fonctionner sur le casque sans latence, et enfin l'articulation étroite entre reconnaissance et adaptation qui caractérise cette approche.

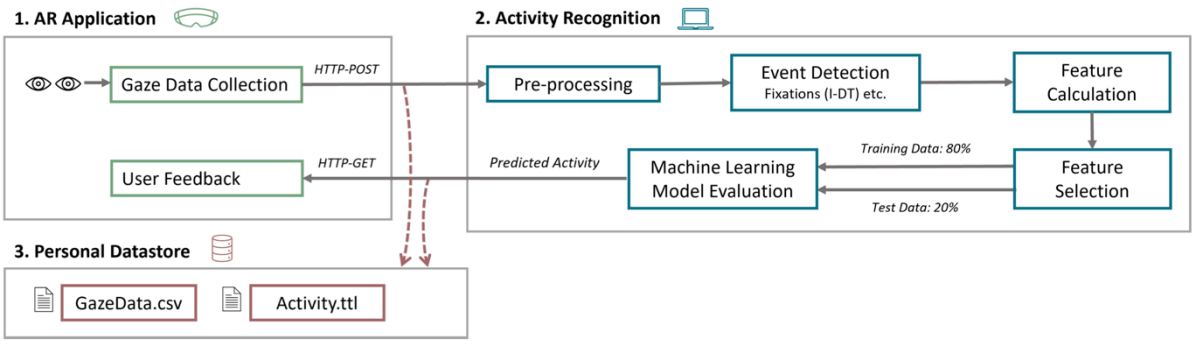


Figure 6. Architecture du système GEAR de reconnaissance d'activité basée sur le regard en AR : de la collecte des données à l'adaptation en temps réel (Bektaş et al., (2024)

Une autre particularité de cette approche se traduit par l'opacité de ses modèles, particulièrement marquée pour les architectures les plus performantes. Cette opacité prend une importance accrue, particulièrement quand la reconnaissance agit sur l'environnement immersif, où une classification erronée produit une adaptation inappropriée. La restitution des critères de décision devient alors un enjeu opératoire autant qu'épistémologique. Explainable XR (Kim et al. 2025) illustre une extension récente du modèle qui adresse ce point. Le système enregistre les interactions de l'utilisateur durant la session et mobilise un large modèle de langage pour produire, à destination du chercheur, des explications en langage naturel des classifications effectuées par le modèle de reconnaissance.

5.3 Discussion

Cette approche apporte une lecture opératoire de l'activité car elle ne se contente pas de représenter ce qui s'est passé mais reconnaît automatiquement ce qui se passe sans nouvelle intervention humaine. Cette propriété ouvre des usages inaccessibles aux deux autres approches, comme la segmentation automatique de larges corpus pour des analyses comparatives ou l'intégration de la reconnaissance dans une boucle adaptative qui ajuste l'environnement au cours du déploiement de l'activité.
Le modèle qu'elle construit présente cependant une limite structurelle. Il ne peut reconnaître que les activités présentes dans son vocabulaire d'entraînement. Toute activité imprévue, ambiguë ou hybride sera forcée dans une catégorie existante ou marquée comme inconnue. Cette fermeture est constitutive de l'apprentissage supervisé et ne se résout pas par un travail méthodologique, seul un réentraînement avec un vocabulaire élargi permet de la dépasser. Elle conditionne ainsi le domaine d'applicabilité de l'approche aux situations où les activités d'intérêt peuvent être définies de manière stable en amont.
La mise en œuvre de cette approche soulève par ailleurs deux points de vigilance. Le premier point concerne la qualité de la reconnaissance qui dépend directement de celle de l'étiquetage initial du corpus d'entraînement. Sa généralisation à de nouvelles populations, modalités ou environnements physiques demande une attention particulière où les variations sont considérables et peuvent dégrader la fiabilité d'un modèle entraîné dans des conditions initiales plus stables. Le deuxième point est relatif à l'explicabilité qui devient critique lorsque la reconnaissance agit sur l'environnement. En effet, les architectures les plus performantes sont aussi les plus opaques et le choix d'une architecture engage un compromis entre fiabilité et lisibilité des décisions de reconnaissance et d'adaptation.

6. Discussion transversale

6.1. Vers un guide méthodologique pour la modélisation de l'activité en XR

L'analyse des trois approches ouvre une lecture plus fondamentale qu'une simple comparaison entre alternatives méthodologiques. Cette lecture tient au fait qu'il n'existe pas une seule manière de modéliser l'activité en XR parce que l'activité elle-même n'y apparaît jamais sous une forme unique. Elle peut être saisie tantôt comme un enchaînement d'actions situées, tantôt comme un ensemble de régularités mesurables, tantôt comme un motif reconnaissable à partir de ses traces. Chacune des approches étudiées opère ainsi une réduction propre : (1) le codage rend visibles les articulations fines de l'interaction mais exige un travail d'interprétation situé ; (2) les variables mesurables permettent de comparer et de généraliser mais en reposant sur des choix initiaux de mesure et de formalisation ; (3) la reconnaissance automatise ces opérations mais dépend de ce qui a été appris en amont.

Ces réductions ne sont pas équivalentes car elles définissent le type de connaissance que l'analyse produit. Lorsqu'on cherche à étendre l'analyse à des volumes de données plus importants ou à des situations en temps réel, certaines dimensions de l'activité doivent être simplifiées, formalisées ou déléguées à des modèles computationnels. Lorsqu'on privilégie une compréhension fine et située, l'analyse gagne en richesse mais perd en extensibilité. Ce compromis ne se résout pas indépendamment du projet de recherche, il définit en même temps la nature de la connaissance produite et le rôle du chercheur dans le processus. Ce rôle se transforme à mesure que l'on passe d'une approche à l'autre. Dans le codage, le chercheur reste l'opérateur central, c'est son geste interprétatif qui transforme les traces en modèle. Dans l'approche par variables mesurables, ce rôle se partage avec un dispositif qui calcule en amont et c'est le chercheur qui interprète ensuite les distributions produites. Dans la reconnaissance, c'est le programme entraîné qui prend en charge l'opération elle-même mais sur la base d'un travail d'étiquetage que le chercheur a réalisé en amont.

Le chercheur n'est donc jamais soustrait du processus de modélisation. Son intervention se déplace et la division du travail entre ce qu'il assume directement et ce qu'il confie au dispositif se redessine.
La question méthodologique ne peut donc plus être formulée comme un choix entre approches. Elle consiste à se positionner dans cet espace de transformations, entre ce qu'on cherche à rendre visible, ce qu'on accepte de simplifier, ce qu'on souhaite rendre comparable ou généralisable et la manière dont on accepte de partager le travail de modélisation avec l'outil. Le tableau 1 propose une lecture dans cette perspective. Il ne départage pas les approches mais rend explicites les implications de leur mobilisation.

Il offre ainsi un cadre pour situer un travail de recherche, comprendre ses compromis, et le cas échéant articuler plusieurs approches dans une même démarche de modélisation.

Tableau 1. Cartographie des trois approches de modélisation de l'activité en XR.

Dimensions	**Approche par Codage et annotation**	**Approche par Variables mesurables**	**Approche par Reconnaissance automatique de l'activité**
Type de modèle produit	Modèles séquentiels et interactionnels	Modèle d'indicateurs comportementaux	Modèle de classification d'activité
Nature qualitative ou quantitative	Du qualitatif avec possibilité de quantification	Quantitatif, interprété qualitativement en aval	Combinaison : entrées quantitatives, corpus qualitatif, sortie qualitative
Volume de corpus	Limité à modéré	Modéré à étendu	Étendu
Analyse ex-situ / in-situ	Ex-situ privilégiée ; In-situ avec grille simplifiée	Ex-situ et In-situ	Ex-situ et In-situ
Division du travail entre chercheur et outil	• Codage et annotation par l'humain, • Formalisation et visualisation instrumentées	• Mesure et calcul instrumentés • Interprétation humaine en aval	• Étiquetage humain en amont • Reconnaissance instrumentée en aval

6.2 Hybridation des approches

Comme précisé plus haut, les trois approches ne sont pas mutuellement exclusives et les pipelines d'analyse en XR les combinent fréquemment. Un codage manuel peut produire le corpus d'entraînement d'un modèle de reconnaissance, des indicateurs quantitatifs peuvent caractériser des segments préalablement codés, une reconnaissance automatique peut segmenter un flux qui sera ensuite exploré qualitativement. Ces articulations relèvent d'une composition méthodologique assumée dans laquelle une approche joue le rôle structurant et d'autres interviennent en complément à des étapes spécifiques du pipeline. Certains dispositifs vont jusqu'à intégrer plusieurs approches dans une même interface d'analyse comme ReLive (Hubenschmid et al., 2022) qui combine un rejeu immersif relevant du codage et de l'annotation avec un dashboard de variables relevant des indicateurs comportementaux en synchronisant les deux vues mutuellement.

La possibilité de telles articulations modifie la manière de lire le guide proposé. Le chercheur ne choisit pas une approche une fois pour toutes mais identifie la démarche principale qui structure son projet et examine dans quelle mesure les autres peuvent intervenir à des étapes spécifiques.

6.3 Limites et perspectives transversales

Si chaque approche assume ses propres limites, certaines questions les traversent toutes. Trois d'entre elles méritent d'être soulignées comme directions ouvertes pour les recherches à venir.

Limite épistémologique : le sens vécu de l'activité. Le positionnement de cet article a consisté à tirer parti d'une spécificité majeure de la XR qu'est la captation native de flux continus et multimodaux qui rendent l'activité partiellement traçable. Cette posture, si productive soit-elle, doit reconnaître sa limite. Les trois approches saisissent l'activité par les traces qu'elle laisse sans accéder à l'expérience subjective que l'acteur fait de son activité, ce que la tradition le cours d'action (Theureau, 2006). Ce que l'acteur perçoit, ressent ou anticipe au moment où il agit excède ce que ses traces révèlent. La XR offre pourtant une opportunité peu exploitée. Il s'agit de la rejouabilité immersive de la session permettant ainsi des autoconfrontations où l'acteur revoit la scène exacte qu'il a vécue et verbalise son expérience dans son contexte même. Articuler les approches présentées à des méthodes compréhensives constitue une piste intéressante pour dépasser cette limite phénoménologique.

Limite de granularité : l'activité au-delà de la session. Les trois approches se déploient à l'échelle de la session immersive, typiquement une heure ou deux. Cette focale devient insuffisante si l'on projette l'évolution des dispositifs XR vers des usages plus continus, mobiles et ubiquitaires. La Time Geography (Hägerstrand, 1970) offre ici un cadre pertinent. Elle représente l'activité humaine comme une trajectoire dans un espace-temps continu et soumise à des contraintes de capabilités, de co-présence et d'accessibilité. Appliquée à la XR (Shaw, 2023), elle ouvre la possibilité de penser l'activité immersive non comme un épisode isolé mais comme une pratique inscrite dans le quotidien professionnel, pédagogique ou social à l'image des smartphones.

Limite éthique : les enjeux de protection des données. Les trois approches reposent sur la collecte de traces dont la sensibilité est souvent sous-estimée. Les mouvements de tête permettent une réidentification proche de la biométrie, le regard révèle des états attentionnels et cognitifs et les signaux physiologiques relèvent du domaine biomédical. Leur combinaison produit un profil comportemental d'une richesse inédite qui excède ce que l'utilisateur peut anticiper lorsqu'il consent à participer à une étude. Le cadre réglementaire qui se précise, dont la RGPD et le AI ACT européen, structure ce qu'il est possible de collecter, de partager et la manière dont le consentement doit être construit. Une perspective émergente se développe actuellement à cet égard à travers la notion de consentement dynamique. Celle-ci propose de remplacer le consentement initial unique par un dispositif continu permettant à l'utilisateur d'ajuster, en cours d'une session XR, ce qu'il accepte de partager et selon quel usage.

7. Conclusion

La modélisation de l'activité en contexte XR ne relève pas d'un simple choix d'outil ni d'une optimisation technique. Elle engage une conception implicite de l'activité humaine et une manière spécifique de transformer des traces en représentations structurées. Les environnements XR, par leur instrumentation native et leurs flux multimodaux continus, rendent certaines dimensions de l'activité particulièrement visibles sans pour autant épuiser sa complexité.
L'analyse proposée a permis d'identifier trois logiques distinctes de transformation des traces : le codage et l'annotation multimodale, la modélisation par variables mesurables, et la reconnaissance automatique de l'activité. Ces approches ne se distinguent pas par leur sophistication technique, mais par l'emplacement de l'opération principale de structuration du modèle. Chacune engage des hypothèses spécifiques sur ce qui constitue une unité pertinente d'analyse de l'activité.

La discussion transversale a mis en évidence que ces trois logiques ne se concurrencent pas mais opèrent chacune une réduction propre, qui rend visibles certaines dimensions de l'activité au prix d'en laisser d'autres en retrait. Leur choix engage à la fois la nature de la connaissance produite, le rôle du chercheur dans le processus de modélisation et la division du travail entre l'humain et l'outil.

Ce que cet article propose, au-delà du choix d'une approche, est un cadre de positionnement : modéliser l'activité en XR, c'est moins arbitrer entre méthodes concurrentes que situer son projet dans un espace de transformations en fonction de la question posée, des contraintes du terrain et des ressources mobilisables. À mesure que les dispositifs immersifs gagnent en ubiquité, en mobilité et en intégration au quotidien des utilisateurs, l'enjeu se déplace vers d'autres questions : Quelles activités souhaite-t-on rendre intelligibles ? A quelles échelles temporelles ? Avec quel partage du travail interprétatif entre l'humain et la machine ? Au prix de quelles renonciations méthodologiques ? La typologie présentée ici n'épuise pas ces questions ; elle entend en donner les premiers repères.

Référencement

Arshad, M. H., Bilal, M., & Gani, A. (2022). Human Activity Recognition: Review, taxonomy and open challenges. *Sensors, 22*(17), 6463. https://doi.org/10.3390/s22176463

Bakeman, R., & Gottman, J. M. (1997). *Observing interaction: An introduction to sequential analysis* (2e éd.). Cambridge University Press. https://doi.org/10.1017/CBO9780511527685

Barbier, J.-M. (2013). *Le travail et la formation des adultes*. Presses Universitaires de France.

Bektaş, K., Strecker, J., Mayer, S., & Garcia, K. (2024). Gaze-enabled activity recognition for augmented reality feedback. *Computers & Graphics, 119*, 103909. https://doi.org/10.1016/j.cag.2024.103909

Card, S. K., Moran, T. P., & Newell, A. (1983). *The psychology of human-computer interaction*. Lawrence Erlbaum Associates.

Hägerstrand, T. (1970). What about people in regional science? *Papers of the Regional Science Association, 24*(1), 7–21. https://doi.org/10.1007/BF01936872

Hubenschmid, S., Wieland, J., Fink, D. I., Batch, A., Zagermann, J., Elmqvist, N., & Reiterer, H. (2022). ReLive: Bridging in-situ and ex-situ visual analytics for analyzing mixed reality user studies. *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems* (p. 1–20). ACM. https://doi.org/10.1145/3491102.3517550

Javerliat, C., Villenave, S., Raimbaud, P., & Lavoué, G. (2024). PLUME: Record, replay, analyze and share user behavior in 6DoF XR experiences. *IEEE Transactions on Visualization and Computer Graphics, 30*(5), 2087–2097. https://doi.org/10.1109/TVCG.2024.3372107

Kim, Y., Aamir, Z., Singh, M., Boorboor, S., Mueller, K., & Kaufman, A. E. (2025). Explainable XR: Understanding user behaviors of XR environments using LLM-assisted analytics framework. *IEEE Transactions on Visualization and Computer Graphics, 31*(5), 2756–2766. https://doi.org/10.1109/TVCG.2025.3549537

Lammert, A., Rendle, G., Immohr, F., Neidhardt, A., Brandenburg, K., Raake, A., & Froehlich, B. (2024). Immersive Study Analyzer: Collaborative immersive analysis of recorded social VR studies. *IEEE Transactions on Visualization and Computer Graphics, 30*(11), 7214–7224. https://doi.org/10.1109/TVCG.2024.3456146

Latour, B. (2005). *Reassembling the social: An introduction to actor-network-theory*. Oxford University Press.

Leontiev, A. N. (1978). *Activity, consciousness, and personality*. Prentice-Hall.

McNeill, D. (1992). *Hand and mind: What gestures reveal about thought*. University of Chicago Press.

Milgram, P., & Kishino, F. (1994). A taxonomy of mixed reality visual displays. *IEICE Transactions on Information and Systems, E77-D*(12), 1321–1329.

Nasri, M., Kosa, M., Chukoskie, L., Moghaddam, M., & Harteveld, C. (2024). Exploring eye tracking to detect cognitive load in complex virtual reality training. *2024 IEEE International*

Symposium on Mixed and Augmented Reality Adjunct (ISMAR-Adjunct) (p. 51–54). IEEE. https://doi.org/10.1109/ISMAR-Adjunct64951.2024.00022

Rakkolainen, I., Farooq, A., Kangas, J., Hakulinen, J., Rantala, J., Turunen, M., & Raisamo, R. (2021). Technologies for multimodal interaction in extended reality: A scoping review. *Multimodal Technologies and Interaction, 5*(12), 81. https://doi.org/10.3390/mti5120081

Rozumnyi, D., Bertsch, N., Sbai, O., Arcadu, F., Chen, Y., Sanakoyeu, A., Kumar, M., Herold, C., & Kips, R. (2024). XR-MBT: Multi-modal full body tracking for XR through self-supervision with learned depth point cloud registration. *arXiv*. https://arxiv.org/abs/2411.18377

Shaw, S.-L. (2023). Time geography in a hybrid physical–virtual world. *Journal of Geographical Systems, 25*(3), 339–356. https://doi.org/10.1007/s10109-023-00407-y

Slater, M., & Sanchez-Vives, M. V. (2016). Enhancing our lives with immersive virtual reality. *Frontiers in Robotics and AI, 3*, 74. https://doi.org/10.3389/frobt.2016.00074

Stevens, S. S. (1946). On the theory of scales of measurement. *Science, 103*(2684), 677–680. https://doi.org/10.1126/science.103.2684.677

Suchman, L. A. (1987). *Plans and situated actions: The problem of human-machine communication*. Cambridge University Press.

Syiem, B. V., Türkay, S., Gallagher, C., & Schrank, C. (2026). An epistemic network analysis of communication strategies during drawing-supported spatial dialogue in VR. *International Journal of Human–Computer Studies, 209*, 103725. https://doi.org/10.1016/j.ijhcs.2025.103725

Theureau, J. (2006). *Le cours d'action : méthode développée*. Octarès.

Varela, F. J., Thompson, E., & Rosch, E. (1991). *The embodied mind: Cognitive science and human experience*. MIT Press.

Vygotsky, L. S. (1978). *Mind in society: The development of higher psychological processes*. Harvard University Press.

Wang, Z.-M., Rao, M.-H., Ye, S.-H., Song, W.-T., & Lu, F. (2025). Towards spatial computing: Recent advances in multimodal natural interaction for extended reality headsets. *Frontiers of Computer Science, 19*(12), 1912708. https://doi.org/10.1007/s11704-025-41123-8

Wittenburg, P., Brugman, H., Russel, A., Klassmann, A., & Sloetjes, H. (2006). ELAN: A professional framework for multimodality research. *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC 2006)* (p. 1556–1559). European Language Resources Association.