Capturer et analyser des gestes professionnels situés : pipeline multimodale, défis techniques et outils pour l'analyse de l'activité

Romaric SICHLER, Théo AKBAS, Alina GLUSHKOVA, Sotiris MANITSARIS
Centre de Robotique, Mines Paris, PSL Université

Résumé

En combinant capture du mouvement par capteurs inertiels, vidéos exocentrées et égocentrées, et enregistrements audios, nous proposons un protocole reproductible adapté à la captation de gestes professionnels situés. Nous décrivons les choix techniques liés à la synchronisation, au prétraitement et à l'alignement des données et présentons deux outils logiciels dédiés : l'un pour l'alignement et la segmentation multimodale, l'autre pour l'élicitation experte à partir de supports vidéo. Ce dispositif vise à préserver la richesse temporelle et située du geste professionnel tout en facilitant la constitution de données exploitables pour l'analyse de l'activité et la comparaison expert–novice.

Mots-clefs

Analyse de l'activité, Geste situé, Acquisition multimodale, Synchronisation des données, élicitation vidéo, Dispositifs Technologiques, Granularité Temporelle

1. Introduction – Multimodalité et action située dans l'analyse du geste professionnel

L'analyse du mouvement humain s'est largement développée dans des environnements contrôlés, où les tâches et les conditions expérimentales peuvent être précisément définies (Sigrist et al., 2013; van der Kruk et al., 2018). Les systèmes de capture du mouvement, optiques ou inertiels, fournissent des données cinématiques fines et ont montré leur intérêt pour l'étude de gestes techniques, y compris en contextes industriels ou artisanaux semi-réels (Olivas-Padilla et al., 2021 ; Olivas-Padilla et al., 2023).
Cependant, l'analyse de gestes professionnels situés soulève des difficultés spécifiques. En situation réelle, l'action est étroitement liée à l'environnement, aux outils et aux matériaux. Ces dimensions introduisent une variabilité importante du geste, parfois constitutive de l'expertise, mais difficile à modéliser de manière exhaustive ou par des métriques absolues (Tsuyuki et al., 2022 ; Vandervoorde et al., 2022 ; Manitsaris, 2021). L'enjeu n'est donc pas uniquement d'accroître la quantité de données disponibles, mais de préserver le caractère situé et temporel de l'activité tout en produisant des données exploitables et comparables.
Cette perspective conduit à des dispositifs multimodaux combinant données de mouvement, vidéo et audio. La multimodalité permet de relier la dynamique corporelle au contexte de l'action et d'intégrer la parole et le point de vue de l'opérateur pour accéder à des critères qualitatifs du geste et de sa réalisation (Pouw et al., 2018 ; Vandevoorde et al., 2022). Toutefois, cette richesse ne peut être mobilisée pour l'analyse que si (i) les flux hétérogènes sont synchronisés et alignés de manière robuste, et si (ii) le sens de l'action (ajustements perceptifs, critères de qualité, erreurs typiques) est documenté par des connaissances expertes ancrées temporellement, notamment via des dispositifs d'élicitation. En contexte professionnel non contrôlé, la mise en œuvre in situ de dispositifs multimodaux se heurte à un ensemble de difficultés méthodologiques rarement explicitées de manière systématique, liées notamment à l'hétérogénéité des dispositifs, à l'absence d'horodatage commun, à la dérive temporelle, aux contraintes de portabilité et à la segmentation de l'activité.

Dans cet article, nous adressons la question suivante : comment concevoir un cadre opératoire instrumenté permettant de capturer, synchroniser et exploiter des données multimodales de gestes situés en environnement professionnel non contrôlé, tout en intégrant l'élicitation experte pour soutenir l'analyse de l'activité ? Nos contributions sont ainsi les suivantes : (i) une cartographie des principaux défis méthodologiques liés à la capture multimodale, à la synchronisation, au prétraitement, à l'alignement et à la segmentation ; (ii) une réponse méthodologique à ces défis à travers un pipeline de capture multimodale située combinant données biomécaniques, vidéos exocentrées/égocentrées et audio ; (iii) une réponse outillée avec deux dispositifs logiciels facilitant a) l'alignement, la visualisation et la segmentation traçable ainsi que b) l'élicitation vidéo en indexant et en alignant texte et vidéo (intervalle temporel vidéo et verbalisation experte).

2. Cartographie des défis de la capture multimodale des gestes professionnels situés

Défi (1) de captation située et de non-intrusivité du dispositif : En environnement professionnel, l'action est fortement conditionnée par l'organisation de l'espace, la présence d'outils et de matériaux variés, les déplacements de l'opérateur et les interactions fines avec la matière. Ces conditions du geste situé, rendent difficile le déploiement de dispositifs de laboratoire à la fois lourds, encombrants et intrusifs, dont l'installation, le transport et la mise en œuvre sont peu compatibles avec le travail réel. Elles imposent, de fait, le recours à des dispositifs autonomes, mobiles, et plus tolérants aux occlusions et aux variations de configuration.
Défi (2) de contextualisation - rendre observable ce qui "fait sens" dans l'action : Dans des gestes situés, la cinématique corporelle seule ne suffit pas à documenter l'activité : les effets produits, les interactions avec les outils, la matière et l'espace de travail participent directement au sens du geste. La captation multimodale (mouvement, vidéo, audio) apparaît alors comme une nécessité pour préserver le caractère situé de l'action. Toutefois, cette contextualisation introduit d'emblée une complexité supplémentaire : multiplier les modalités, c'est multiplier les sources d'hétérogénéité (dispositifs, formats, points de vue, qualité de signal) et donc les conditions de comparabilité et d'interprétation.
Défi (3) de synchronisation et d'alignement temporel intermodalités : La capture multimodale mobilise des dispositifs hétérogènes fonctionnant à des fréquences d'échantillonnage et selon des mécanismes d'horodatage distincts. En contexte réel, ces dispositifs sont déployés de manière autonome, sans infrastructure centralisée : une synchronisation fiable devient alors un prérequis pour toute analyse conjointe. Au-delà d'un alignement global, la synchronisation conditionne la possibilité de relier événements visuels, sonores et cinématiques, et donc de préserver le sens du geste tel qu'il se déploie dans l'action. Dans ce cadre, la synchronisation doit généralement être assurée a posteriori, ce qui expose notamment à des risques de dérive temporelle et à des écarts d'alignement difficiles à détecter sans outils adaptés.
Défi (4) de traitement des données - bruit, hétérogénéité des formats : Les données acquises en situation réelle sont susceptibles de contenir du bruit. Par ailleurs, certaines modalités (notamment les données biomécaniques) peuvent être représentées par des formats hiérarchiques, avec des métadonnées temporelles incomplètes ou non directement compatibles avec les outils multimédias classiques. Ces caractéristiques rendent le prétraitement, l'alignement et la vérification de cohérence plus coûteux, et augmentent le risque d'erreurs silencieuses (par exemple, un alignement temporel incorrect qui reste plausible visuellement).
Défi (5) de segmentation - définir des unités analytiques comparables : L'exploitation analytique des données multimodales suppose de segmenter l'activité afin d'isoler des portions pertinentes (phases, actions, primitives) et de produire des segments cohérents et comparables entre enregistrements. En captation située, la segmentation est rendue plus complexe par la variabilité des conditions d'exécution et par la nécessité d'articuler plusieurs modalités sur une même échelle temporelle.

Défi (6) d'interprétabilité et d'accès au savoir expert - l'élicitation indexée dans le temps : Dans les métiers manuels, une part importante de l'expertise repose sur des ajustements fins liés à la perception, à l'expérience corporelle et à l'interaction avec la matière, qui échappent aux descriptions normatives du geste. Le visionnage de vidéos de sa propre pratique permet à l'expert de se replacer dans la situation d'action et de verbaliser des éléments du geste qui resteraient autrement implicites (Vermersch, 2019). Toutefois, une difficulté méthodologique majeure apparaît dès lors que l'on souhaite exploiter ces verbalisations : associer explicitement une transcription textuelle à des intervalles temporels précis de la vidéo, condition nécessaire à des recherches multimodales fiables et à des usages ultérieurs (comparaison expert–novice, exploitation pédagogique).
Ces défis soulignent la nécessité d'outils adaptés pour manipuler conjointement des données multimodales hétérogènes : import, visualisation sur une même échelle temporelle, contrôle de cohérence, découpage/segmentation.

3. Réponse méthodologique : pipeline de capture et préparation des données multimodales

3.1 Instrumentation multimodale in situ

Pour documenter la cinématique du geste et son contexte d'exécution, nous proposons un pipeline de capture multimodal : données biomécaniques (motrice), vidéo et audio. Pour la modalité motrice, nous utilisons une combinaison et des gants inertiels (*Nansense Biomed Bundle*, 50 capteurs) ; les données sont représentées sous forme de squelette articulé et encodées au format BioVision Hierarchy (BVH). Ce choix vise à rendre la captation compatible avec les contraintes de terrain (Défi 1) (mobilité, occlusions) tout en conservant une représentation exploitable de la cinématique corporelle. Les capteurs inertiels permettent d'accéder à des indicateurs de coordination, de posture et d'ergonomie ; malgré leur caractère invasif et leur sensibilité aux perturbations magnétiques, ils constituent, dans notre expérience, un compromis adapté pour l'enregistrement de gestes situés.
La vidéo est mobilisée pour contextualiser l'action (Défi 2) (outils, matière, espace de travail, effets produits) et servir de support aux analyses et à l'élicitation. Nous combinons une vue exocentrée (organisation globale, interactions opérateur–outils–objets) et une vue égocentrée (zone d'action, contacts, champ de vue). Dans notre protocole, nous privilégions des caméras compactes portées sur la tête (*GoPro Hero 12*) afin de maintenir une captation moins intrusive qu'une caméra lourde ou embarquée dans des lunettes, (la caméra peut être équipée sur un casque de sécurité ou dans un boitier ATEX) tout en conservant une information visuelle directement liée au geste.
Les données audios complètent la description de l'activité en capturant l'ambiance de l'atelier (microphones à électret) et, lorsque nécessaire, des signaux de contact (capteurs piézoélectriques). Elles documentent des événements significatifs (contacts, chocs, frottements), difficiles à inférer des seules données visuelles ou cinématiques. Nos acquisitions ont été effectuées avec un *Zoom H4e*, compatible avec une capture stéréo et une synchronisation par code temporel Bluetooth.
Ce pipeline a déjà été mobilisé par notre équipe pour l'enregistrement de gestes artisanaux dans plusieurs métiers : soufflage de verre à la canne, verrerie scientifique, joaillerie, argenterie, cultivation du mastic, tissage de la soie, ganterie, sellerie maritime, taille du marbre et porcelaine. Cette diversité de terrains ne permet pas de conclure à une transférabilité générale du dispositif, mais elle constitue un premier indice de sa portée méthodologique au-delà d'un seul cas d'étude. Une partie de ces données est déjà disponible[1]. Les autres seront ajoutés au fur et à mesure de l'obtention des droits de diffusion et de leur traitement, afin d'enrichir un corpus multimodal de gestes professionnels situés exploitable pour l'analyse (Olivas-Padilla et al., 2023), la comparaison et, à terme, des usages pédagogiques.

[1] https://www.caor.minesparis.psl.eu/human-motion-capture-benchmark/
Une partie de la capture de ces données a été soutenu par le Gouvernement français dans le cadre du projet « ReSource » (AMI Compétences et métiers d'avenir), relevant du programme France 2030 et opéré par la Caisse des Dépôts.

3.2 Traitement des données : synchronisation, filtrage, alignement et segmentation

La synchronisation des modalités est assurée *a posteriori* (Défi 3) selon deux approches. La première repose sur une synchronisation matérielle par code temporel audio (*générateurs Deity TC-1*), enregistrée sur les pistes audio/vidéo, robuste sur la durée mais plus lourde. La seconde, utilisée dans la majorité des enregistrements, consiste à assurer un horodatage commun au démarrage puis à aligner les flux ; la durée des tâches observées (≈ dix minutes) limite l'impact de la dérive temporelle et rend cette stratégie suffisamment précise (dérive maximale : 120ms) pour la plupart de nos usages (reconnaissance de mouvement, collaboration humain machine). Concrètement, les caméras sont déclenchées via QR codes *(GoPro Labs*), l'enregistreur audio est synchronisé via Bluetooth (*Zoom H4e/BTA-1*), et les données inertielles héritent de l'horloge système de la machine d'acquisition. Le fabricant recommande une resynchronisation afin de limiter la dérive temporelle pour les enregistrements de plus d'une heure. Dans notre cas, pour des acquisitions allant jusqu'à quinze minutes, nous n'avons pas observé de désynchronisation supérieure à 20 ms (résolution des caméras), ce qui reste compatible avec nos besoins analytiques (segmentation en primitive du geste, analyse croisée entre données expertes et apprenties).
Un prétraitement est ensuite appliqué afin de réduire le bruit et corriger les artefacts en préservant la traçabilité temporelle (Défi 4) : filtrage passe-bas et corrections manuelles si nécessaire (*MotionBuilder*) pour le mouvement, nettoyage audio coupe bas. Les modalités sont ensuite rééchantillonnées afin d'être projetées sur une échelle temporelle commune (sec. 4), définie à partir de la modalité la plus lente (dans notre cas la vidéo) condition nécessaire à l'exploitation conjointe des événements multimodaux (sec. 5).
Enfin, une segmentation temporelle semi-automatique isole des portions pertinentes de l'activité (gestes, phases de tâche : actions, primitives) et produit des segments cohérents et comparables entre enregistrements. La mise en œuvre opérationnelle de l'alignement et de la segmentation multimodale, ainsi que la gestion de formats hétérogènes, sont détaillées dans la section suivante.
La dernière étape du pipeline consiste à éliciter les séquences vidéo afin de documenter le sens de l'action et produire une annotation sémantique (Défi 6). Les participants commentent ce qu'ils observent et font, fournissant une donnée complémentaire mobilisable pour caractériser les gestes. Pour être exploitable en analyse multimodale, cette verbalisation doit être indexée temporellement : la création de paires segment-texte alignées sur des intervalles vidéo précis constitue un prérequis méthodologique, motivant le dispositif présenté en section 5.

4. Réponse outillée : dispositif d'alignement, visualisation et segmentation traçable

La capture multimodale de gestes professionnels repose sur des dispositifs et des formats hétérogènes, dont certains (notamment les formats de mouvement hiérarchiques BVH/TC) sont peu compatibles avec les outils multimédias génériques (*Premiere Pro, ANVIL, etc)*, rendant l'alignement et la segmentation coûteux, sources d'erreurs difficiles à détecter. Nous proposons un outil visant une interopérabilité temporelle minimale, import multimodal, construction d'une timeline commune, contrôle de cohérence et découpage synchronisé, afin de soutenir la reproductibilité du traitement des données. Un retour d'usage interne suggère une réduction d'environ 60 % du temps d'alignement et de segmentation par rapport à des outils génériques de traitement multimédia ; une évaluation plus systématique reste néanmoins nécessaire pour en apprécier la robustesse et la généralité.[2]

[2] L'outil est disponible ici : https://github.com/RSichler/multimodal_alignement_tool

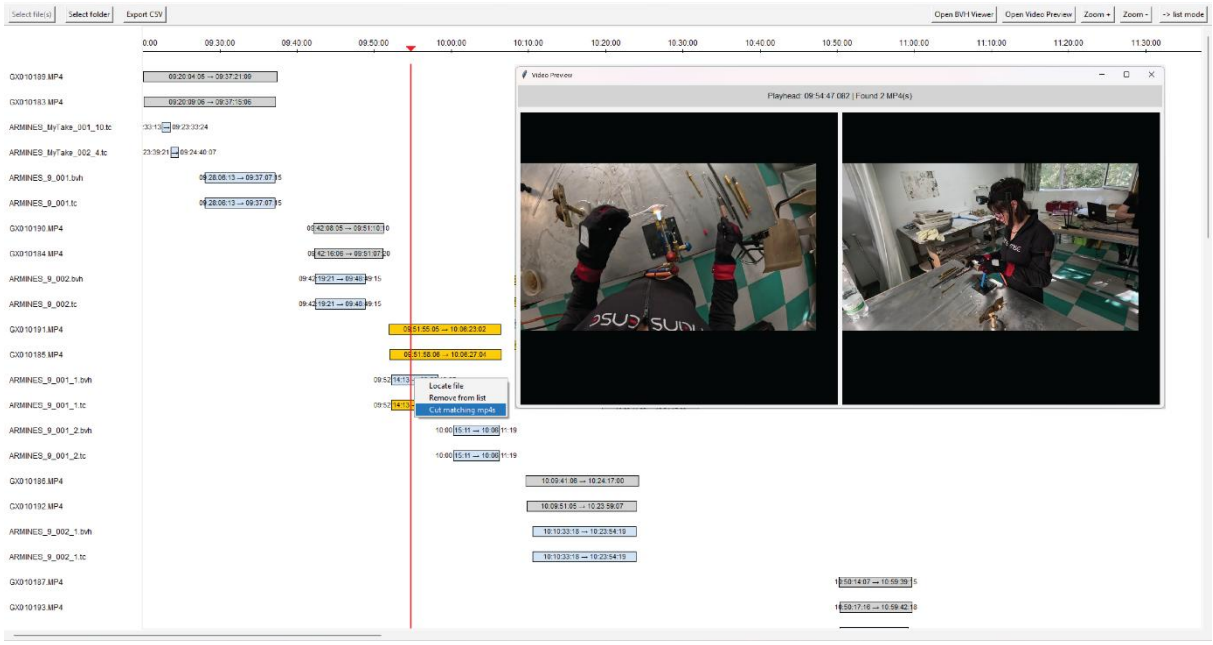

Fig. 1 – Vue de l'outil d'alignement, de visualisation et de segmentation multimodale

4.1 Interopérabilité temporelle multimodale : import, extraction temporelle et timeline commune

L'outil permet d'importer des fichiers wav, mp4, bvh et tc. À l'import, il recherche les informations temporelles disponibles y compris lorsque l'information temporelle est encodée sur un canal audio, afin d'établir pour chaque flux une référence temporelle exploitable. Une difficulté spécifique concerne les formats de mouvement : les fichiers BVH décrivent une temporalité relative (échantillons référencés au premier échantillon) sans horodatage absolu, alors que le format TC exporté par le logiciel du fabricant intègre l'information temporelle. L'outil met en correspondance ces fichiers afin d'associer une information temporelle aux données de mouvement hiérarchiques lorsque celle-ci est disponible via TC, et de faire coexister mouvement, audio et vidéo sur une timeline commune.

4.2 Visualisation pour le contrôle de cohérence et segmentation

À partir de cette timeline, l'outil propose une vue temporelle conjointe des flux (fig. 1), conçue comme un support de contrôle de cohérence de l'alignement (repérage d'événements, détection d'écarts intermodalités) avant l'analyse. Il permet ensuite de découper les flux de manière synchronisée : les segments produits partagent des bornes temporelles communes, de sorte que l'instant *t* d'une modalité corresponde à celui des autres. Une visualisation primaire des données facilite le repérage d'événements et soutient la segmentation ; ce découpage synchronisé vise à rendre la segmentation traçable et reproductible. Enfin, l'outil est conçu pour pouvoir être adapté à d'autres formats comportant une information temporelle, ce qui ouvre la possibilité d'un usage dans d'autres métiers, sous réserve d'ajustements liés à l'activité, aux environnements de travail et aux instruments mobilisés.

5. Réponse outillé : Dispositif d'élicitation experte indexée sur la vidéo

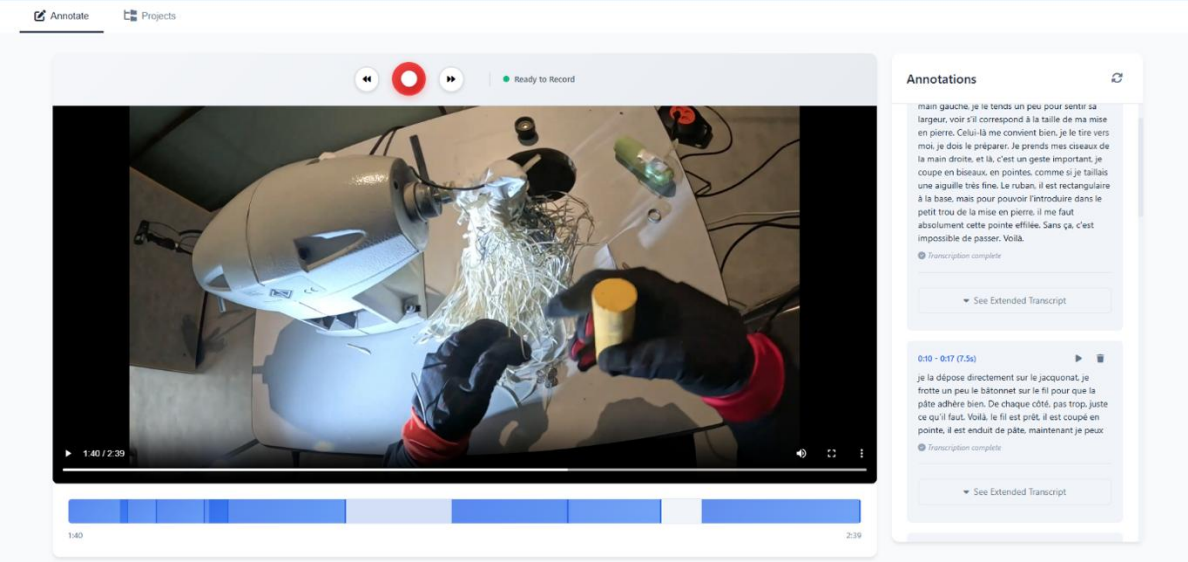

Fig. 2 – Dispositif d'élicitation de vidéo pour la constitution de savoir expert aligné vidéo/parole

Dans les métiers manuels, une part importante de l'expertise repose sur des ajustements fins liés à la perception, à l'expérience corporelle et à l'interaction avec la matière, difficilement formalisables par des descriptions normatives du geste. L'élicitation à partir de supports vidéo constitue un cadre méthodologique pertinent pour articuler discours expert, contexte visuel et temporalité de l'action (Vermersch, 2019). Dans une approche multimodale, l'enjeu est de rendre ces verbalisations exploitables en les indexant temporellement et en les reliant explicitement à des segments vidéo**.** Pour répondre à cet enjeu, nous proposons un dispositif d'élicitation instrumenté dont la contribution centrale est la production d'une unité de donnée explicite et réutilisable : un intervalle vidéo associé à une verbalisation experte. L'expert sélectionne un segment vidéo et enregistre un commentaire vocal attaché à cet intervalle ; celui-ci est transcrit automatiquement puis affiché sous une forme éditable, garantissant un lien direct entre texte et temporalité vidéo.

5.1 Implémentation

Le dispositif est implémenté sous la forme d'une application web client–serveur (fig. 2). L'interface permet la lecture vidéo, la sélection d'intervalles et l'enregistrement vocal associé, ainsi que l'affichage et l'édition des transcriptions. Le backend (FastAPI, Python) assure la persistance des annotations (SQLite) et orchestre la chaîne de traitement : transcription automatique (Whisper-v3-turbo via Fireworks.ai, langue française), puis génération de contenus dérivés (transcription enrichie, étiquettes catégorisées) à l'aide d'un modèle de langage interrogé par API (Llama-v3.3-70B-Instruct). Les traitements sont exécutés de manière asynchrone, avec retour d'état en temps réel vers l'interface (WebSocket), afin de préserver la fluidité de l'annotation. L'implémentation maintient une séparation explicite entre données primaires (intervalle vidéo, audio, transcription éditée) et productions dérivées, garantissant traçabilité et exploitabilité ultérieure.[3]

5.2 Alignement temporel texte/vidéo et exploitabilité pour l'analyse multimodale

Chaque annotation est structurée autour de l'association explicite entre un intervalle temporel vidéo et une verbalisation experte. Elle comprend les bornes temporelles du segment, l'enregistrement audio associé et une transcription textuelle éditable, constituant la donnée experte primaire exploitable pour l'analyse. Cette structuration garantit l'ancrage des verbalisations dans des instants précis de l'action et leur articulation avec les autres modalités synchronisées du corpus.

[3] L'outil est disponible ici : https://github.com/Somekindofa/video_elicitation_annotation_tool

En complément, le système génère des productions dérivées destinées à l'indexation et à l'exploration : transcriptions enrichies (description de l'exécution, erreurs fréquentes, conseils correctifs) et étiquettes catégorisées (outil, matériau, technique, manipulation), tout en maintenant une distinction claire avec les données primaires.
L'ensemble des annotations est exportable au format JSON structuré (intervalles, transcriptions, marqueurs, métadonnées), fournissant une base directement exploitable pour des analyses ultérieures et garantissant la traçabilité et la réutilisabilité du corpus.

6. Conclusions

Cette proposition ne constitue pas une solution technique universellement transférable à l'ensemble des métiers manuels, mais une réponse méthodologique à une difficulté centrale de l'analyse de gestes situés : articuler captation, synchronisation et exploitation de données multimodales pour la documentation et l'analyse automatique. Sa transférabilité repose moins sur les équipements eux-mêmes que sur les principes qu'il organise : instrumentation, alignement temporel, segmentation traçable et élicitation indexée, dont la mise en œuvre doit être ajustée aux contraintes propres à chaque terrain.
Nous proposons une réponse outillée à une difficulté centrale de l'analyse de l'activité en contexte réel : constituer des corpus multimodaux de gestes professionnels à la fois situés, exploitables et comparables. En articulant une cartographie explicite des défis rencontrés in situ, un pipeline de capture et de préparation des données, et deux outils logiciels dédiés à l'alignement multimodal et à l'élicitation experte indexée, nous proposons une approche pour préserver le sens temporel et contextuel du geste tout en améliorant la traçabilité et la reproductibilité des traitements.
Les outils présentés permettent notamment de réduire significativement le coût temporel de la préparation des données et de transformer des verbalisations expertes en annotations multimodales directement exploitables pour l'analyse, la comparaison expert–novice ou des usages pédagogiques.
Plusieurs limites demeurent néanmoins. Le protocole est assez spécifique à notre équipement qui est coûteux et repose toujours sur des opérations partiellement manuelles (correction, segmentation, annotation). Les perspectives de recherche portent sur l'évaluation systématique des outils, l'automatisation prudente de certaines étapes (segmentation, détection d'événements, aide à l'annotation) pour permettre de traiter davantage de données, plus rapidement et l'extension de l'approche à d'autres domaines professionnels et dispositifs de capture.

Références

- Sigrist, R., Rauter, G., Riener, R., & Wolf, P. (2013). Augmented visual, auditory, haptic, and multimodal feedback in motor learning: A review. *Psychonomic Bulletin & Review*, 20(1), 21–53. https://doi.org/10.3758/s13423-012-0333-8
- van der Kruk, E., & Reijne, M. M. (2018). Accuracy of human motion capture systems for sport applications; state-of-the-art review. *European Journal of Sport Science*, 18(6), 806–819. https://doi.org/10.1080/17461391.2018.1463397
- Olivas-Padilla, B., Manitsaris, S., Menychtas, D., & Glushkova, A. (2021). Stochastic biomechanic modeling and recognition of human movement primitives in industry using wearables. *Sensors*, 21(7), 2497. https://doi.org/10.3390/s21072497
- Olivas-Padilla, B., Glushkova, A., & Manitsaris, S. (2023). Motion capture benchmark of real industrial tasks and traditional crafts for human movement analysis. *IEEE Access*, 11, 40075–40092. https://doi.org/10.1109/ACCESS.2023.3269581
- Tsuyuki, S., Hoshina, K., Miyahara, K., Suhara, M., Matsukura, M., Isaji, T., & Takayama, T. (2022). Motion analysis of suturing technique with Leap Motion Controller™: Proof-of-concept. *Science Progress*, 105(3). https://doi.org/10.1177/00368504221103777
- Pouw, W., Trujillo, J. P., & Dixon, J. A. (2018). The quantification of gesture–speech synchrony: A tutorial and validation of multimodal data acquisition using device-based and video-based motion tracking. *Behavior Research Methods*, 52, 723–740. https://doi.org/10.3758/s13428-019-01271-9
- Vandevoorde, K., De Meester, S., De Clercq, D., & De Pauw, K. (2022). Using artificial intelligence for assistance systems to bring motor learning principles into real-world motor tasks. *Sensors*, 22(7), 2481. https://doi.org/10.3390/s22072481
- Manitsaris, S. (2021). Movement-based Human-Machine Collaboration: a Human-centred AI approach (accreditation to supervise research) (Doctoral dissertation, Sorbonne Université).
- Vermersch, P. (2019). L'entretien d'explicitation. ESF Sciences humaines. https://doi.org/10.3917/esfsh.verme.2019.01