Analyse, reconnaissance et synthèse d'expressions et de styles dans les mouvements

Alexandre Meyer

Université Claude Bernard Lyon 1

Année 2023 N° XX–2023

Habilitation à Diriger Des Recherches
en
Informatique

présentée et soutenue publiquement le 30 mars 2023 par

Alexandre MEYER

Analyse, reconnaissance et synthèse d'expressions et de styles dans les mouvements

Laboratoire LIRIS - UMR 5205 CNRS
École Doctorale "Informatique et Mathématiques" (ED512)
Spécialité Informatique

COMPOSITION DU JURY

Pr.	Boubakeur BOUFAMA	Professeur, University of Windsor, Canada	Rapporteur
Pr.	Marie-Paule CANI	Professeure, École Polytechnique	Rapporteure
Pr.	Renaud SEGUIER	Professeur, Centrale Supelec	Rapporteur
Pr.	Saïda BOUAKAZ	Professeure, Université Claude Bernard Lyon 1	Examinatrice
D.R.	Edmond BOYER	Directeur de recherche, INRIA	Examinateur
Pr.	Jean-Claude MARTIN	Professeur, Université Paris Saclay	Examinateur
D.R.	Catherine PELACHAUD	Directrice de recherche, CNRS	Examinatrice

Résumé

Les émotions sont un des aspects importants de la vie quotidienne. Elles s'expriment par différents canaux, dont certains sont visuels comme les expressions du visage, les postures et les mouvements du corps. En complément de la parole, il s'agit là d'un moyen de communication important qui permet de fournir ou de capter des informations sur l'état émotionnel d'une personne. En informatique, les applications cherchant à prendre en compte l'état émotionnel de l'utilisateur ou cherchant à produire des personnages virtuels expressifs sont de plus en plus nombreuses : l'éducation, la médecine, le commerce, la sécurité, la robotique et les loisirs (cinéma, jeux vidéo, etc.). Dans leurs interactions homme-machine, ces applications essaient de tendre vers des interactions gestuelles et conversationnelles où les expressions prennent une place importante. En synthèse d'images, les personnages virtuels animés sont présents partout dans les médias vidéo tels que les films, les jeux vidéo ou les mondes virtuels. La façon dont ces personnages se déplacent, agissent dans leur monde et interagissent est un point d'attention important pour leurs créateurs. Et pour les rendre vivants et attractifs, il faut les doter de gestes réalistes et d'émotions. Pour répondre à cette attente, les systèmes doivent être capables d'analyser et de comprendre les comportements, c'est-à-dire l'état émotionnel, d'une personne, mais ils doivent aussi être capables de produire ou d'aider à la production d'une animation de haute qualité, expressive et diversifiée. Nos travaux s'inscrivent dans cette optique de comprendre, reconnaître et synthétiser des gestes, des mouvements, des animations du visage et du corps produisant une expression ou un style. La notion d'expression et de style est le fil conducteur de nos travaux en vision par ordinateur et en synthèse d'animations. Nous proposons des avancées en analyse et reconnaissance d'expressions faciales et corporelles ; ainsi qu'en synthèse et édition d'animations comportant un style ou une expression particulière.

Mots-clés : reconnaissance d'expressions faciales, reconnaissance d'expressions corporelles, animation procédurale, animations basée sur des données.

Abstract

Emotions are an important aspect of daily life. They are expressed through different channels, some of which are visual : facial expressions, postures and body movements. In addition to speech, this is an important means of communication that can provide or capture information about a person's emotional state. In computer science, applications seeking to take into account the emotional state of the user or seeking to produce expressive virtual characters are increasingly numerous : education, medicine, commerce, security, robotics and entertainment (cinema, video games, etc.). In their human-machine interactions, these applications try to move towards gestural and conversational interactions where expressions take an important place. In image synthesis, animated virtual characters are present everywhere in video media such as movies, video games or virtual worlds. The way these characters move, act in their world and interact is an important focus for their creators. And to make them lively and attractive, they need realistic gestures and emotions. To meet this expectation, systems must be able to analyze and understand the behaviors, i.e. the emotional state, of a person, but they must also be able to produce or assist in the production of high quality, expressive and diverse animation. Our work takes place in this perspective of understanding, recognizing and synthesizing gestures, movements, animations of the face and body producing an expression or a style. The notion of expression and style is the common thread of our work in computer vision and animation synthesis. We offer advances in facial and body expression analysis and recognition, as well as in animation synthesis and editing with a particular style or expression.

Keywords : facial expression recognition, body expression recognition, procedural animation, data-driven animation.

Table des matières

1 Introduction

Table des matières du chapitre

Les émotions sont un des aspects importants de notre vie quotidienne. Les émotions transparaissent par différents canaux plus ou moins perceptibles par les autres, comme la parole, les textes, les signaux physiologiques, les expressions du visage, les postures et les mouvements du corps. Il s'agit là d'un moyen de communication important qui permet de fournir ou de capter des informations sur l'état émotionnel d'une personne afin de mieux comprendre ou transmettre ses intentions et son ressenti. Dans les expressions non verbales, on a longtemps pensé que les expressions faciales fournissaient la majorité des informations sur les émotions ressenties. Cependant, des travaux dans le domaine de la psychologie ont montré que les expressions corporelles fournissaient autant, voire plus d'informations sur l'état émotionnel d'une personne. Ceci montre l'intérêt de considérer les deux types d'expressions afin de pouvoir fournir des systèmes capables de montrer ou de décoder l'état émotionnel d'un individu. De plus, même si l'expression d'une émotion se transmet par plusieurs canaux, la prise en compte d'une sous-partie de ces canaux comme le visage ou la posture reste un moyen important d'améliorer les connaissances. Généralement, les travaux traitant simultanément tous les canaux combinent des approches spécifiques dédiées à un seul canal. Comme une personne émet des expressions et perçoit celles des autres, il paraît primordial d'aborder les expressions visuelles par les deux aspects allant de l'analyse, la perception, la reconnaissance jusqu'à la manière dont elles s'expriment.

Les relations entre les mouvements du corps et les expressions ont d'abord été principalement étudiées par les Sciences Humaines et Sociales. La psychologie cherche à comprendre quels facteurs induisent quels sentiments afin de mieux appréhender les interactions humaines. Les arts visuels ont également tenté de formaliser la relation entre le geste et l'émotion afin d'améliorer les spectacles vivants par le jeu des acteurs ou des danseurs humains, mais également en proposant des animations de personnages virtuels plus réalistes, car dotés d'émotions. En informatique, les applications cherchant à prendre en compte l'état émotionnel de l'utilisateur ou cherchant à produire des personnages virtuels expressifs sont de plus en plus nombreuses dans les domaines de l'éducation, la médecine, le commerce, la sécurité, la robotique, les loisirs (cinéma, jeux vidéo), etc. Pour les relations entre l'humain et la machine ou entre les personnages virtuels, ces applications essaient de tendre vers des interactions gestuelles et conversationnelles où les expressions prennent une place importante. Il a été montré que des systèmes informatiques ou robotiques plus sensibles à ce que ressent l'utilisateur permettent d'améliorer l'acceptation de la machine. Donner une sorte d'empathie à une machine est le prochain grand défi des systèmes informatiques.

En synthèse d'images, les personnages virtuels animés sont omniprésents dans les médias vidéo tels que les films, les films d'animation ou les jeux vidéo. On les retrouve souvent sous les feux de la rampe, racontant des histoires ou incarnant des joueurs, mais aussi en arrière-plan, peuplant des mondes virtuels. Ils renforcent l'immersion des spectateurs en rendant les mondes virtuels plausibles. La façon dont ces personnages se déplacent et agissent entre eux et avec l'environnement est un point d'attention important pour leurs créateurs. Pour les rendre vivants et attractifs, il faut les doter d'émotions. Un personnage virtuel expressif ou un robot empathique et expressif sont plus facilement acceptés, car assimilés à des êtres vivants. Pour faire face à ce défi, les systèmes doivent être capables d'analyser et de comprendre les comportements, c'est-à-dire l'état émotionnel d'une personne, mais

ils doivent aussi être capables de produire ou d'aider à la production d'une animation de haute qualité, expressive et diversifiée.

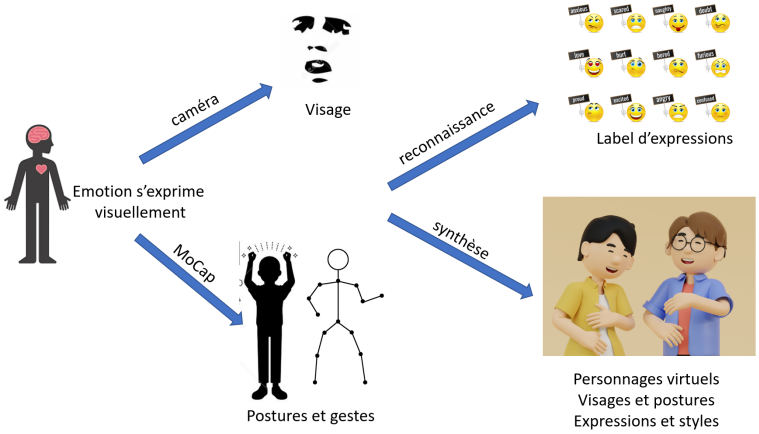


Figure 1.1 : Les émotions s'expriment par de nombreux canaux. Nous ne nous intéressons qu'aux expressions faciales et corporelles en cherchant à en effectuer la reconnaissance ou la synthèse. En synthèse d'images, nous considérons la notion de style au sens large qui englobe tous les qualificatifs qui peuvent décrire un geste.

Nos travaux s'inscrivent dans cette optique de comprendre, reconnaître et synthétiser des gestes, des mouvements, des animations du visage ou du corps produisant une expression ou un style (voir la figure 1.1). Les expressions sont liées à l'émotion, alors que le terme style, souvent utilisé en synthèse, est plus large, incluant tous les adjectifs qui peuvent qualifier un geste comme ceux liés à la vitesse, à la démarche, à l'âge de la personne, à ses caractéristiques physiques, etc. Les notions d'expressions et de style sont le fil conducteur de nos travaux. Nous cherchons à reconnaître les expressions sur les visages, dans les postures et les mouvements du corps. Nous traitons également des expressions sans les catégoriser, mais en les capturant directement sur un visage depuis des images afin de les transférer sur un modèle 3D. Nous décomposons les expressions en caractéristiques compréhensibles et explicables pour pouvoir les reconnaître et les éditer.

1.1 De la vision à la synthèse et réciproquement

Pour animer un personnage virtuel exprimant différentes expressions ou styles, il faut d'abord observer le phénomène, l'analyser et le comprendre. La tâche la plus abordable pour être sûr d'avoir appréhendé les bonnes caractéristiques est sûrement de commencer par produire une méthode de reconnaissance automatique telle que la réalise la communauté de vision par ordinateur. D'un autre côté, le domaine de la synthèse d'animations a l'habitude de travailler sur des données de mouvements humains : des textures, des maillages 3D, des squelettes en mouvement, etc. Les outils de synthèse peuvent donc aider à analyser et à reconnaître des expressions. Il semble naturel de travailler sur ces deux aspects. Même si la reconnaissance, l'analyse et la synthèse d'expressions sont adressées par plusieurs communautés distinctes avec leurs codes et leurs habitudes : la vision par ordinateur pour la reconnaissance automatique d'expressions, la synthèse d'images pour la création, ainsi

que la communauté d'Interaction Homme Machine pour les travaux sur les émotions dans les interactions.

Expressions faciales

L'état de l'art en vision et en synthèse montre que l'essentiel des travaux sur les expressions s'est concentré sur le visage. La reconnaissance d'expressions faciales est un sujet qui a été largement étudié, mais souvent dans un environnement contrôlé : expressions actées et non spontanées, visage fixe face à la caméra, lumière statique, sujets adultes, etc. Pour espérer être efficace dans toutes les conditions, il reste de nombreux verrous à aborder. Il faut traiter chaque spécificité méthodiquement. Dans ce cadre, nous abordons le cas d'images à faibles résolutions et le cas des expressions de visages d'enfants, deux spécificités rarement étudiées précédemment.

Pour animer un visage en 3D, le domaine se base beaucoup sur la capture de mouvements. Des dispositifs capturant des visages de manières très réalistes existent, mais ils se basent sur plusieurs dizaines de caméras, complétées par des centaines de sources de lumière. En revanche, il y a peu de travaux visant à démocratiser la capture de mouvements du visage en se basant sur un environnement extrêmement simple avec une unique caméra. Pourtant, de tels systèmes, moins coûteux et plus mobiles, pourraient servir à de nombreuses applications pour produire des animations où une qualité moindre serait suffisante. En mêlant vision et animation, nous proposons des techniques de capture de déformations de peau (rides) à partir d'une unique caméra, afin de les transférer à un avatar virtuel.

Expressions corporelles

Les études menées dans le domaine de la psychologie ont montré que l'expression corporelle est aussi pertinente que les expressions faciales pour exprimer les émotions. Parallèlement, avec la démocratisation croissante des dispositifs de capture de mouvements, il est aussi abordable de travailler sur les mouvements du squelette qu'avec des images de visage. Les chercheurs en reconnaissance automatique ont beaucoup travaillé sur la reconnaissance d'actions, mais bien moins sur la reconnaissance d'expressions. En synthèse, la communauté s'est d'abord beaucoup concentrée sur la capacité d'animer des personnages effectuant des actions de base comme la marche, la course, les changements de direction, etc. Il reste un champ de recherche important autour des styles et des expressions.

La communauté de vision cherche des méthodes automatiques de reconnaissance en mettant au point des caractéristiques discriminantes. En synthèse d'animations, il est important de trouver des approches permettant d'offrir à un animateur des capacités d'édition de caractéristiques compréhensibles. Identifier et formaliser des caractéristiques pertinentes et explicables est un des verrous clés que nous abordons. À partir d'une méta-analyse qualitative et quantitative des publications en psychologie, IHM, vision par ordinateur et synthèse d'images, il est possible d'en obtenir une liste. Pour les valider, nous proposons de les utiliser dans un processus de reconnaissance automatique d'expressions corporelles. Le deuxième grand défi que nous relevons est de trouver des outils intuitifs d'éditions de ces caractéristiques. Il faut proposer à l'utilisateur mieux qu'un travail fasti-

dieux sur chaque articulation de chaque pose et mieux qu'une technique tout automatique sans contrôle. Nous proposons à l'animateur de pouvoir exprimer son savoir-faire en éditant ces caractéristiques par des outils utilisant l'apprentissage profond ou l'animation procédurale.

Les données

Dans de nombreux domaines de l'informatique, les avancées très récentes sont souvent issues de l'apprentissage qui exploite automatiquement des données. En particulier, les réseaux de neurones tirent leur puissance dans leur capacité à extraire des généralités dans les données, qui doivent alors être suffisamment importantes et représentatives. Pour le domaine très spécifique abordé ici (la reconnaissance et la synthèse des expressions du visage et du corps) des données existent, mais les très grands ensembles de données sont souvent difficiles à obtenir. Dans l'idéal, il faudrait des expressions du visage, des expressions corporelles, produites par un grand nombre de sujets différents, voire des animaux différents, dans des situations différentes, de préférence spontanées, et ceci avec une grande précision de capture. Dans la réalité, il faut souvent se contenter de données qui ne répondent que partiellement à ces critères. Il est donc important de ne pas faire reposer tous les travaux de recherche sur la disponibilité de ces données.

À l'extrême, nous pouvons citer les techniques procédurales qui se basent sur des assemblages de petites procédures expertes qui résolvent chacune une sous-partie d'un problème et n'utilisent aucune donnée. Dans certains cas comme l'animation de créatures virtuelles imitant les insectes ou ayant des morphologies imaginaires, la piste semble intéressante, car aucune donnée n'existe. Pour de nombreux autres problèmes en reconnaissance et en animation de mouvements du corps des données existent, mais en quantité et en représentativité limitées. Les approches couplant ces données aux connaissances d'un expert sont alors pertinentes pour lever de nombreux verrous. La majorité de nos travaux s'inscrit dans ce cas-là. Même lorsque nous utilisons les réseaux de neurones, nous veillons à ce que le temps d'entraînement reste raisonnable, et que les données soient simples à acquérir. Nous argumentons qu'une voie de recherche intéressante est de contraindre les réseaux de neurones à être plus explicables en forçant dans leur fonction de coût à identifier des critères compréhensibles par un humain. Par exemple, pour l'édition de poses, nous proposons d'ajouter des réseaux capables de modifier une représentation d'un squelette dans l'espace latent abstrait que produisent les réseaux en respectant des caractéristiques utilisables par un animateur. Cette voie est à poursuivre pour de nombreux problèmes en synthèse d'animations afin de rendre les réseaux compatibles avec la manière de travailler des artistes qui veulent pouvoir intervenir dans le processus.

1.2 Organisation du manuscrit

Après un chapitre 2 dédié à l'état de l'art, le chapitre 3 présente des contributions liées à la reconnaissance d'expressions faciales. Une première contribution vise à reconnaître les expressions du visage en s'inspirant d'une étude perceptive du système visuel humain. La deuxième contribution porte sur une étude des expressions spontanées de visages d'enfants en proposant une nouvelle base de données, car les enfants sont peu présents

dans les corpus de données. Nous proposons une approche de reconnaissance d'expressions spécialisées sur les visages d'enfant par transfert d'apprentissage. Le chapitre 4 présente des contributions liées à l'analyse et à la reconnaissance d'expressions posturales. Il présente quatre grandes familles de techniques pour réaliser cette tâche. Ces familles sont comparées avec comme critère d'évaluation le taux de bonne reconnaissance, mais également par la capacité à expliquer quelles caractéristiques sont importantes. En effet, ces caractéristiques doivent servir pour proposer des techniques de synthèse d'expressions dans le chapitre 5. Ce chapitre propose des approches liées à la capture et à l'animation de visages, des techniques d'animation procédurale de créatures à n pattes et deux familles de méthodes pour la synthèse de gestes expressifs. Enfin, la conclusion dresse un bilan et ouvre des perspectives de poursuites de nos travaux.

1.3 Co-encadrements de thèses et projets

La grande majorité des travaux présentés dans ce manuscrit ont été réalisés dans le cadre de co-encadrements de thèses dont la liste est ci-dessous. Un grand merci à tous les collaborateurs pour leurs implications dans ces travaux. Pour de plus amples détails sur chaque partie, il est possible de se reporter au manuscrit correspondant. Pour chaque chapitre, la liste de nos publications concernées est donnée.

- [Dut11] Ludovic Dutreve. "Paramétrisation et Transfert d'Animations Faciales 3D à partir de Séquences Vidéo : vers des Applications en Temps Réel." Theses. Université Claude Bernard Lyon 1, LIRIS, 2011
  Encadrement : Saïda Bouakaz, Alexandre Meyer.
  Voir le chapitre 5.
- [Kha13] Rizwan A. Khan. "Expression recognition from videos in uncontrolled environment". Theses. Université Claude Bernard Lyon 1, LIRIS, 2013
  Encadrement LIRIS : Saïda Bouakaz, Alexandre Meyer. Encadrement LHC : Hubert Konik.
  Voir le chapitre 3.
- [Abd14] Ahmad Abdul Karim. "Procedural Locomotion of Multi-Legged Characters in Complex Dynamic Environments : Real-Time Applications". Theses. Université Claude Bernard Lyon 1, LIRIS, 2014
  Encadrement LIRIS : Saïda Bouakaz, Alexandre Meyer.
  Encadrement Spirops : Thibaut Gaudin, Axel Buendia. Voir le chapitre 5.
- [Cre19] Arthur Crenn. "Capture et transfert d'expression de visages d'enfants pour l'interaction avec des mondes virtuels". Theses. Université Claude Bernard Lyon 1, LIRIS, 2019
  Encadrement LIRIS : Saïda Bouakaz, Alexandre Meyer. Encadrement LHC : Hubert Konik.
  Voir les chapitres 3 et 4.
- [Vic23] Léon Victor. "Learning-Based Interactive Character Animation : Expressing Emotions through Motion". Theses. INSA de Lyon, LIRIS, 2023
  Encadrement : Saïda Bouakaz, Alexandre Meyer.
  Voir le chapitre 5.

- [Mah23] Mehdi-Antoine Mahfoudi. “Génération procédurale d’animations porteuses d’expressions : approche temps réel et interactive”. Theses. Université Claude Bernard Lyon 1, LIRIS, 2023
  Encadrement LIRIS : Saïda Bouakaz, Alexandre Meyer. Encadrement Spirops : Thibaut Gaudin, Axel Buendia.
  Voir les chapitres 4 et 5.

Les collaborations, projets et financements liés à ces travaux sont les suivants.

- 2007-2009 FUI PlayAll. Projet réunissant 9 acteurs industriels du monde du jeu vidéo et 5 laboratoires. L’objectif était le développement d’un moteur de jeux associé à des outils d’aide à leur création. La thèse de Ludovic Dutreve [Dut11] a fait partie de ce projet.
- 2009-2011 FUI PlayAll Online. La suite du projet PlayAll.
- 2009-2012 projet CNRS/entreprise. Collaboration entre le *LIRIS* et l’entreprise *Spirops* pour la thèse d’Ahmad Abdul Karim [Abd14] autour de l’animation procédurale de créatures à n pattes.
- 2010-2013 Projet région. Collaboration entre le *LIRIS* et le laboratoire Hubert Curien de Saint-Étienne autour de la thèse de Rizwan Khan [Kha13].
- 2015-2018 FUI KurioEye. Projet réunissant KD Interactive, Solidanim, l’Université de Poitiers et le *LIRIS* visant à développer des outils visuels pour l’interaction entre les enfants et une tablette tactile qui leur est spécialement dédiée.
- 2015-2019 Projet région. Collaboration entre le *LIRIS* avec le laboratoire Hubert Curien à Saint-Étienne autour de la thèse d’Arthur Crenn [Cre19].
- 2018-2019 Projet Pulsalys (incubateur Univ. de Lyon). Collaboration entre le *LIRIS* et le Laboratoire Hubert Curien de Saint-Étienne. Financement de 6 mois de contrat ingénieur pour des travaux sur les micro-expressions du visage.
- 2019-2022 Bourse du ministère pour la thèse de Léon Victor [Vic23].
- 2020-2023 CIFRE. Collaboration entre le *LIRIS* et l’entreprise *Spirops* pour la thèse de Mehdi-Antoine Mahfoudi [Mah23] autour de l’animation procédurale expressive.
- Différents projets internes aux *LIRIS* et de la Fédération Informatique de Lyon pour des financements de stage de Master.

2 État de l'art

Table des matières du chapitre

2.1 Communication non verbale

La communication verbale est le moyen prépondérant pour exprimer ses émotions, mais il est maintenant largement admis que les comportements non verbaux [BP93 ; ST11] constituent un autre moyen important de communication, en complément de la parole (voir la figure 2.1). Les expressions faciales offrent une information primordiale aux interlocuteurs, mais les études du domaine montrent que les postures et les gestes du corps dans leur ensemble sont également un moyen indispensable de communication de l'état émotionnel d'une personne [Bui+14].

Dans le monde du numérique, les systèmes d'IHM deviennent de plus en plus élaborés, car ils peuvent se baser de plus en plus sur les mouvements humains capturés grâce aux avancées dans le domaine de la vision. Les applications comme les jeux vidéo, la création artistique, la réalité virtuelle ou les systèmes d'apprentissage du geste bénéficient naturellement des capacités à analyser, comprendre et reproduire un geste. Dans ce contexte, l'étude des moyens de communication non verbale véhiculant principalement les émotions humaines visibles sous forme d'une expression est cruciale. L'analyse d'expressions fournit de nouvelles informations pour obtenir des interactions plus naturelles. De plus, la généralisation du numérique permet une acceptation par les utilisateurs de ces outils. Par exemple, *Facebook*, *Microsoft* ou *Google* utilisent la détection des visages et la reconnaissance des expressions faciales pour adapter le contenu proposé. Longtemps considérés comme intrusifs, ces systèmes commencent à se démocratiser et à être acceptés par les utilisateurs. Avant d'aller plus loin, il semble nécessaire d'essayer de différencier ce qu'est un sentiment, une émotion, une expression et un style, puis de répertorier les représentations les plus courantes des émotions.

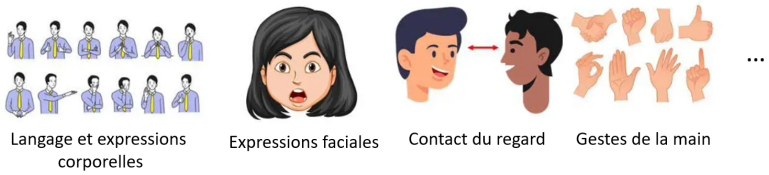


Figure 2.1 : La communication sans utiliser de mots comme le langage corporel, les expressions faciales et les gestes de la main est appelée communication non verbale.

2.1.1 Sentiment, émotion, expression

Définir précisément ce qu'est une émotion n'est pas simple. Bien que tout le monde ressente et exprime des émotions quotidiennement, lorsqu'on demande une caractérisation concrète, personne n'est capable de répondre. Ou plutôt chaque domaine propose une définition qui lui est propre, allant de la définition intuitive du grand public à la définition physiologique en passant par la définition imagée des artistes ou les définitions affectives, cognitives, comportementales, etc. Le Petit Robert offre la définition suivante : *"État affectif intense, caractérisé par une brusque perturbation physique et mentale où sont abolies, en présence de certaines excitations ou représentations très vives, les réactions appropriées d'adaptation à l'évènement. Au sens affaibli : État affectif, plaisir ou douleur, nettement prononcé."* Kleinginna *et al.* [KK81] recensent 92 définitions différentes d'une émotion et tentent d'en extraire

différentes caractéristiques communes en vue de proposer la définition suivante : *les émotions sont le résultat de l'interaction de facteurs subjectifs et objectifs, réalisés par des systèmes neuronaux ou endocriniens, qui peuvent :*

- *induire des expériences telles que des sentiments d'éveil, de plaisir ou de déplaisir ;*
- *générer des processus cognitifs tels que des réorientations pertinentes sur le plan perceptif, des évaluations, des étiquetages ;*
- *activer des ajustements physiologiques globaux ;*
- *induire des comportements qui sont, le plus souvent, dirigés vers un but adaptatif.*

Il est souvent souligné qu'une émotion est une réaction physiologique de notre corps face à un évènement extérieur [Jam84]. Cependant, on peut aussi ressentir une émotion lorsque l'on repense à un évènement passé. Cette diversité de point de vue montre à quel point il est difficile de converger sur une définition précise d'une émotion. Dans une étude basée sur la collecte de descriptions de situations affectives auprès de 1242 personnes, Scherer *et al.* [Sch+04] tentent de caractériser un état affectif selon différents facteurs :

- intensité de l'état affectif ;
- durée de l'état affectif ;
- synchronisation des sous-systèmes de l'organisme ;
- focalisation sur l'évènement ;
- rapidité du changement d'état ;
- impact comportemental.

Dans ces travaux, les chercheurs définissent cinq sous-systèmes : l'évaluation cognitive, les changements psychophysiologiques, l'expression motrice, les tendances à l'action et le sentiment subjectif. Un état affectif va donc être caractérisé par ces cinq composants ainsi que par les différents facteurs présentés ci-dessus. Ils considèrent les émotions comme un sous-ensemble particulier parmi les différents types d'états affectifs. Celles-ci sont plus intenses, mais de plus courtes durées. En particulier, les émotions sont caractérisées par un haut degré de synchronisation entre les différents sous-systèmes. De plus, elles sont susceptibles d'être très axées sur les évènements déclencheurs et produits par l'évaluation cognitive. Du point de vue de leur effet, elles ont un fort impact sur le comportement et sont capables de changer rapidement.

Le neurologue américain Damasio [Dam02] écrit : *Les émotions et les sentiments d'émotions sont respectivement le début et le terme d'une progression, mais le caractère relativement public des émotions et l'aspect complètement privé des sentiments qui en découlent montrent bien que les mécanismes situés tout au long de ce continu sont extrêmement différents.* Il propose de *" réserver le terme sentiment à l'expérience mentale et privée d'une émotion, et d'utiliser au contraire le terme émotion pour désigner l'ensemble de réponses qui, pour bon nombre d'entre elles, sont publiquement observables."* Dans ce manuscrit, le terme *expression* représente la sous-catégorie des réponses émotives observables visuellement, sous-entendu par une ou plusieurs caméras. Une expression est ce qu'une personne peut percevoir sur un visage, une posture du corps, de manière statique ou dynamique. Une expression est donc un moyen de communication non verbale, c'est une manifestation visuelle d'une émotion.

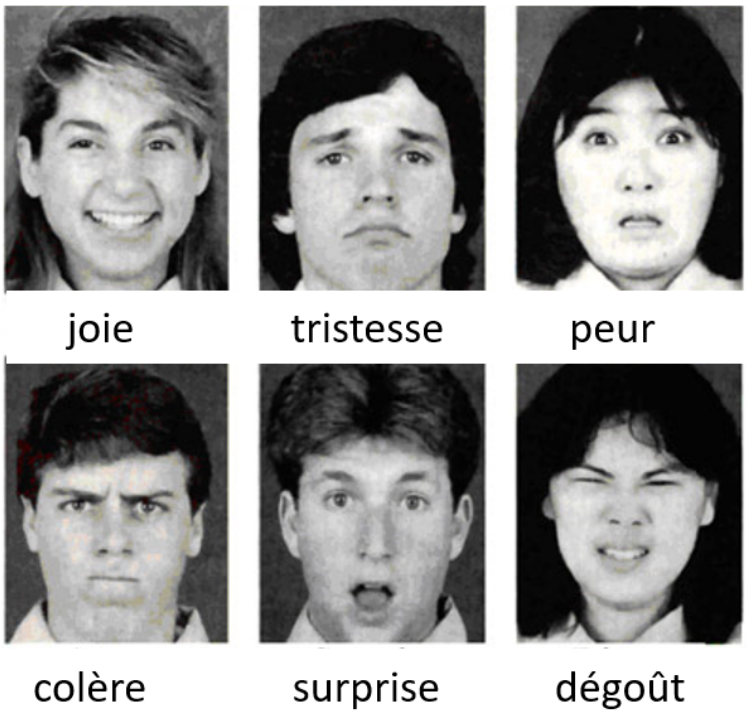


Figure 2.2 : Les six expressions de base définies par Ekman sont présentées sur des visages.

2.1.2 Style

Dans le contexte des mouvements humains, une expression est une qualité qui exprime de manière spontanée une émotion et qui se distingue de l'action que veut faire la personne en parallèle. La notion d'expression est la manifestation visuelle d'une émotion. La notion de style a un sens plus large. L'action correspond à ce qu'une personne cherche à faire, le style au sens large englobe toutes les informations décrivant comment le geste est effectué. L'expression est liée à l'émotion ressentie et retranscrite dans le geste. Le style englobe tous les autres qualificatifs qui peuvent qualifier la manière de faire le geste. Ces qualificatifs peuvent être associés à la morphologie (par ex. âge, jeune, avec une jambe de bois), à la physiologie (par ex. fatigué, sportif, mourant), à l'expérience passée (par ex. maladroit, précis, confus), au contexte de l'environnement ou de la situation (par ex. pressé, perdu, discret, concentré), etc. Cette notion de style est également directement liée au qualificatif décrivant la démarche dans le cas particulier où l'action est la locomotion.

Il est possible d'effectuer un rapprochement avec la notion de style dans la littérature ou les arts visuels. Une définition du Robert sur la notion de style en littérature est la suivante : *"Part de l'expression qui est laissée à la liberté de chacun, n'est pas directement imposée par les normes, les règles de l'usage, de la langue"*. Deux écrivains peuvent raconter une même histoire, mais avec un style différent afin de transmettre des informations complémentaires différentes. Les lois de la physique et de la biomécanique dans un geste sont comme la grammaire dans un texte. Elles définissent des règles à respecter. L'action réalisée (marcher, courir, s'asseoir, utiliser un outil, etc.) correspond au contenu, à l'histoire, à la trame. La notion de style correspond à tous les qualificatifs que l'on peut ajouter sur comment

l'action se déroule. Pour les arts visuels, le Robert donne cette définition : *"Manière de traiter la matière et les formes dans une œuvre d'art."* Le domaine de l'édition de style en traitement d'images est un domaine actif depuis des années [Jin+19]. Ce domaine utilise les notions de contenu et de style. Le contenu fait référence aux formes, le style correspond à la matière, à la couleur et à la texture. Il semble pertinent d'utiliser le même terme style pour parler de ce qui caractérise une animation en complément du contenu qui sera l'action réalisée.

2.1.3 Représentation des émotions

Parmi les styles les plus étudiés en animation, reconnaissance ou synthèse, il y a les émotions. Il est donc important de répertorier les différentes représentations. Le passage du vocabulaire descriptif à des notions qualitatives a déjà été étudié dans le domaine de la psychologie des émotions. Seuls les modèles les plus connus sont présentés dans ce document, ceux utilisés par la communauté "image" au sens large.

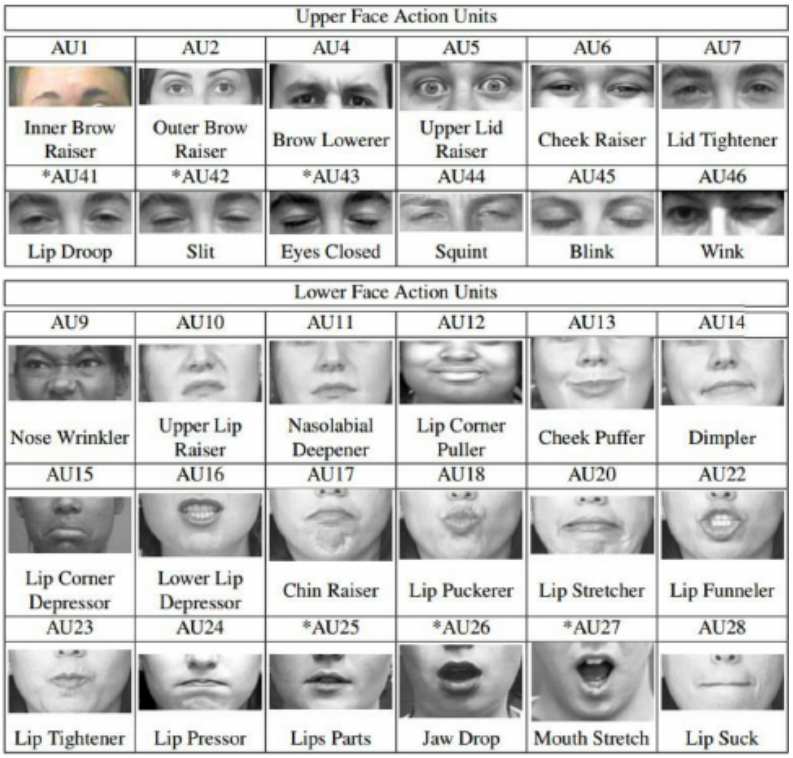


Figure 2.3 : Différentes unités d'action (*AU*) pour la partie haute et basse du visage. Source : [De +15]

Le modèle discret d'Ekman

Le système de codage des expressions faciales et corporelles le plus fréquemment utilisé dans la communauté graphique au sens large est celui proposé par Ekman [Ekm+87]. Ce modèle propose six expressions basiques : la joie, la colère, la tristesse, la peur, le dégoût

et la surprise (voir la figure 2.2). La théorie d'Ekman a été inspirée des travaux réalisés par Darwin [TD77]. Selon Darwin, l'expression des émotions a évolué chez l'homme à partir d'animaux. Darwin soutenait que les expressions n'étaient pas apprises, mais innées dans la nature humaine et qu'elles étaient donc importantes pour la survie en raison de l'évolution. Ainsi, Darwin a réussi à démontrer l'universalité des émotions de base. Cela veut dire que tous les individus du monde, quel que soit leur peuple d'appartenance, ressentent les émotions de bases de la même manière. Comme on peut le voir sur cette image, une simple photographie peut parfois suffire à reconnaître l'expression. Ces six émotions de base sont largement utilisées dans la communauté graphique.

Ekman et Friesen [EF76] proposent un système de codage d'action faciale, *Facial Action Coding System - FACS)* dans le but de décrire les mouvements du visage. Dans ce système, les contractions ou décontractions de zones du visage sont décomposées en unités d'action (*Action Unit - AU*). Le système *FACS* repose sur la description de 46 *AU* identifiées par un numéro dans la nomenclature FACS. Ainsi, l'*AU* 1 va correspondre à l'action de lever les sourcils. Une expression faciale correspond, donc, à la mise en jeu de plusieurs *AU*. Par exemple, dans le cas d'un visage apeuré, les *AU* mises en jeu sont : lever les sourcils intérieurs, lever les sourcils extérieurs, tension des lèvres, abaissement de la mâchoire inférieure. La figure 2.3 représente différentes unités d'action pour la partie haute et basse du visage.

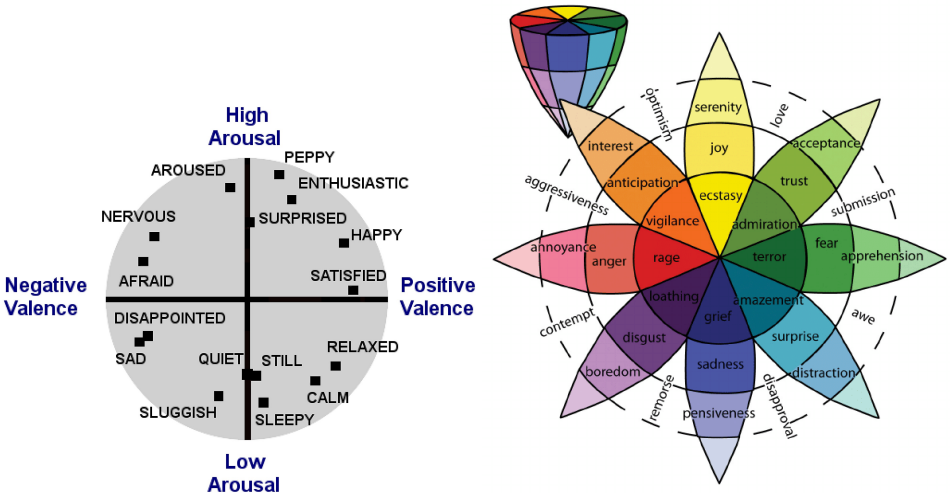


Figure 2.4 : À gauche : espace des émotions arousal - valence de Russel. À droite : la roue des émotions de Plutchik peut être vue comme un modèle hybride entre le modèle purement discret d'Ekman et le modèle continu arousal/valence.

Le modèle arousal - valence

Dans de nombreuses applications, la représentation discrète d'Ekman semble limitée, car elle ne permet pas de nuancer une expression. Plusieurs auteurs dans différents travaux [Rus80 ; CB94] proposent un modèle liant les expressions les unes aux autres. Le modèle le plus abouti, proposé par James A. Russel, repose sur un système 2D à deux axes bipolaires : la valence en abscisse et l'arousal (ou excitation) en ordonnée comme illustré sur la figure 2.4 à gauche. La valence va de la tristesse au bonheur, tandis que l'excitation va de l'ennui

ou de la somnolence à l'excitation frénétique. Selon Russell, les émotions se situent dans un cercle de cet espace bidimensionnel, et sont caractérisées par des catégories floues regroupées dans ce plan 2D.

Bien qu'un espace continu puisse représenter toutes les expressions possibles, cela n'est pas garanti par la théorie de Russell et n'a d'ailleurs pas été démontré. Le passage d'un label discret à une position dans le plan et inversement est souvent subjectif, ce qui est problématique lorsque l'on cherche à reconnaître de manière automatique des expressions. Il est donc peu utilisé en vision par ordinateur, mais sa continuité est un avantage en animation pour permettre des interpolations.

La roue des émotions de Plutchik

Une autre représentation significative des émotions a été proposée par Plutchik [Plu80]. Dans son modèle, les émotions humaines sont modélisées selon huit émotions primaires. Les émotions sont organisées en paires d'opposés : la joie contre la tristesse, la confiance contre le dégoût, la peur contre la colère et l'anticipation contre la surprise (voir la figure 2.4 à droite). Ce modèle peut être considéré comme un modèle hybride entre le modèle d'Ekman et le modèle arousal/valence. Ainsi, les émotions les plus complexes sont définies comme une combinaison d'émotions basiques. Par exemple, l'amour est considéré comme la combinaison de la joie et de la confiance. Dans l'ensemble de nos travaux, les six émotions basiques d'Ekman sont souvent privilégiées. Elles sont plus faciles à utiliser dans les algorithmes de classification, mais réfléchir à utiliser des modèles continus est une voie à explorer, notamment pour les aspects synthèses.

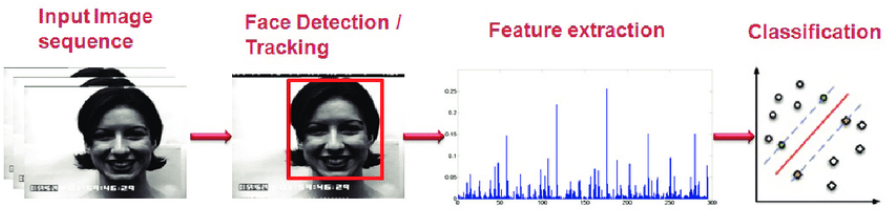


Figure 2.5 : Les étapes "classiques" de la reconnaissance d'expressions faciales sont la localisation du visage, l'extraction des caractéristiques et la classification donnant le label de l'expression.

Le modèle *OCEAN*

Les cinq grands traits de personnalité, souvent appelés OCEAN [Wig96], et parfois CANOE, sont : Ouverture, Conscience, Extraversion, Agréabilité et Névrose. Ces cinq traits représentent de vastes domaines du comportement humain et expliquent les différences de personnalité et de prise de décision. Aujourd'hui, le modèle est utilisé par exemple en ressources humaines ou en marketing pour évaluer les réactions des personnes.

Le lien entre ce modèle et le mouvement est peu souvent abordé. Il est possible de citer Durupinat *et al.* [Dur+09] qui étudient la relation entre la perception des mouvements de foule et la personnalité. Neff *et al.* [Nef+10] évaluent l'effet du geste et du langage sur la perception de la personnalité des personnages virtuels avec lesquels un humain converse.

Puis, Durupinat *et al.* [Dur+17] proposent d'utiliser les caractéristiques de Laban (voir la section 2.3.1) pour modifier la personnalité dans les mouvements humains. À notre connaissance, il n'existe pas de bases de données de mouvements liés à ce modèle.

2.2 Reconnaissance d'expressions faciales

L'étude des expressions faciales à partir d'images ou de vidéos a suscité beaucoup d'intérêt, un documentaire grand public sur l'Intelligence Artificielle en parle comme une illustration majeure du domaine [TV22]. Il serait difficile de faire une étude exhaustive sur ce thème. Le processus classique de reconnaissance automatique d'expressions peut être résumé par les étapes suivantes. La première étape consiste à détecter ou suivre soit le visage, soit le corps, afin d'analyser uniquement la partie pertinente de l'image. Après avoir localisé la région d'intérêt, l'étape suivante consiste à extraire des informations significatives ou discriminantes à travers la formalisation de différents descripteurs [SBW17]. Ces derniers sont ensuite fournis à un classifieur, qui est entraîné sur différentes bases de données, afin de détecter l'expression faciale ou corporelle fournie en entrée du système. La figure 2.5 présente le principe général de détection d'expressions faciales. Le processus pour les expressions corporelles est assez semblable, mais les descripteurs sont différents. L'arrivée des méthodes à base de réseaux de neurones profonds a un peu modifié ce schéma en proposant parfois d'incorporer toutes les étapes d'extraction des descripteurs dans le réseau, famille d'approches qui sera détaillée dans la section 2.2.2. Le manuscrit de thèse de Rizwan A. Khan [Kha13], puis celui d'Arthur Crenn [Cre19] proposent deux états de l'art des travaux existants avant 2019. L'état de l'art de Li *et al.* de 2022 [LD22] et celui de Zago *et al.* [Can+22] donnent une vue d'ensemble du domaine incluant des travaux très récents. Dans ce manuscrit, un survol des grandes familles existantes avant l'apprentissage profond est réalisé avant de présenter les avancées récentes.

2.2.1 Approches dites "classiques"

Il est courant de classer les descripteurs d'expressions du visage en cinq familles : les descripteurs de texture, ceux basés sur la détection de contours, les descripteurs géométriques, les descripteurs globaux et les descripteurs basés sur l'application de *patchs*.

Les descripteurs de texture

Les descripteurs de texture vont chercher à capturer à la fois les caractéristiques des éléments principaux de visage (nez, yeux, bouche), mais également les rides et mouvements de peau qui ont leur importance [CBM09]. Un descripteur classique de texture est le filtre de Gabor. Il inclut des informations de magnitude et de phase. La fonction de magnitude du filtre de Gabor contient les informations sur l'organisation de l'image du visage. La phase est utilisée pour contrôler la façon dont le filtre réagit aux différentes orientations de l'image [BV08 ; OZM14 ; ZTC14 ; Her+16 ; HSH16]. La figure 2.6 présente le résultat de l'application d'un filtre de Gabor sur un visage. On peut observer les différents descripteurs qui vont être obtenus par convolution du filtre sur l'image originale. Les motifs binaires locaux (*Local Binary Pattern - LBP*) [ZP09] sont aussi utilisés pour l'extraction de descripteurs de

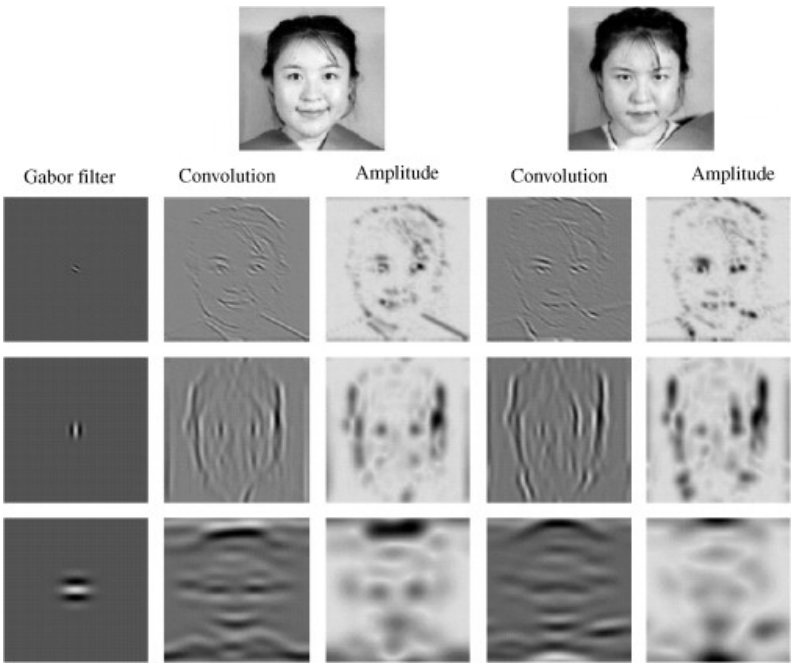


Figure 2.6 : Application d'un filtre de Gabor pour la détection d'expressions faciales. Le filtre est tourné selon plusieurs directions (phase) afin d'avoir différentes réponses et donc produire différentes textures orientées. Source : [BV08]

texture. Le principe général est de comparer le niveau de luminance d'un pixel avec les niveaux de ses voisins. Cela permet d'avoir une information relative à des motifs réguliers dans l'image, autrement dit une texture. Selon l'échelle du voisinage utilisé, certaines zones d'intérêts tels que des coins ou des bords peuvent être détectées par ce descripteur [HR15 ; CNK16]. De nombreuses variantes de descripteurs de texture ont été proposées et sont présentées dans l'état de l'art de Li *et al.* [LD22].

Détection de contours

Les contours des éléments d'un visage donnent de nombreuses informations pour reconnaître l'expression comme illustré sur la figure 2.7. Les *Active Shape Model* [Le+12] sont un outil classique en vision par ordinateur de détection et de suivi de contours d'un objet. Ce modèle est largement utilisé et dérivé pour reconnaître des expressions faciales. Par exemple, Song *et al.* [Son+10] proposent les *Graphics-processing unit based Active Shape Model* où le calcul et la mise en correspondance sont accélérés par le *GPU*. Un autre exemple de descripteur de contours est l'Histogramme des Gradients Orientés (*HOG*) qui calcule des histogrammes locaux de l'orientation du gradient sur des sous-parties de l'image [DTS06]. Dahmane et Meunier [DM14] utilisent les *HOG* pour extraire des caractéristiques visuelles, par exemple une expression de joie est transcrite par une courbure au niveau des yeux. De nombreuses extensions à ces approches existent, notamment récemment en multi-résolutions [NSM18].

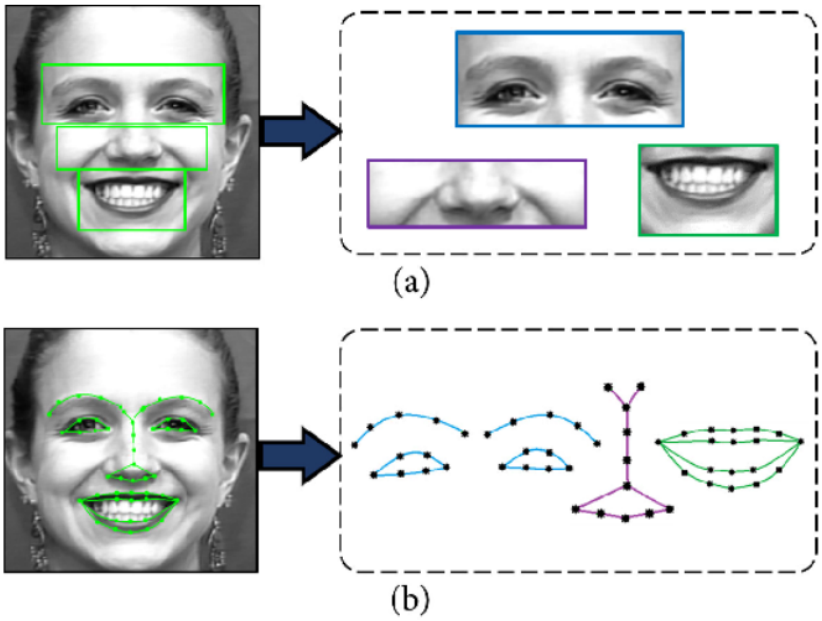


Figure 2.7 : Deux exemples de familles de descripteurs pour la reconnaissance d'expressions faciales. En haut, des *patchs* sont appliqués sur les régions du visage : leur activation ou non donne des indications sur l'expression. En bas, les contours détectés peuvent être donné à un classifieur. Source : [BNK18]

Les descripteurs géométriques

Les descripteurs géométriques cherchent à localiser et à suivre des points caractéristiques sur un visage en espérant être plus précis que les détecteurs de contours vus précédemment [Wan+18a]. Pour cela, certaines méthodes se basent par exemple sur la *Local Curvelet Transform (LCT)* qui est une généralisation en dimension supérieure de la transformée en ondelettes. La *LCT* est conçue pour représenter des images à différentes échelles et sous différents angles. Demir [Dem+14] utilise la moyenne, l'entropie et l'écart type de la *LCT* pour construire un ensemble de descripteurs. Le problème clé avec les méthodes géométriques est de savoir localiser précisément les points caractéristiques et de les suivre sans introduire de bruit. Dans les applications réelles, en raison de la variation de pose, d'éclairement, d'occultations et du bruit provenant de l'entrée, il est très difficile de localiser avec précision ces points avec les approches classiques. Les approches à base d'apprentissage profond présentées dans la section 2.2.2 améliorent considérablement les résultats. Il est à noter que ces descripteurs seuls, tout comme ceux de contours, n'offrent qu'une information partielle.

Descripteurs locaux et globaux

Une approche classique en analyse de données est l'Analyse en Composantes Principales (*ACP*) pour extraire différents descripteurs. En effet, l'*ACP* permet d'extraire des caractéristiques globales des visages avec un nombre de dimensions réduit. Les vecteurs propres d'une base de données d'images de visages exprimant une émotion servent à construire

un espace de toutes les expressions [MFM14]. Les vecteurs propres d'une nouvelle image de visage donnent une combinaison linéaire d'expressions et peuvent être utilisés comme descripteurs. L'analyse en composantes indépendantes [Tay14] ou une analyse discriminante linéaire pas à pas peuvent également être utilisées [Sid+15]. Ces approches sont à mettre en relation avec les approches récentes présentées dans la section 2.2.2 qui se basent sur des auto-encodeurs [Zen+18 ; Hu+22] pour construire un espace latent. Le code de l'espace latent cherche à représenter de manière efficace l'espace des visages, à l'instar des vecteurs propres.

Descripteurs basés sur des *patchs*

L'idée intuitive des *patchs* est d'avoir un calque générique représentant ce que l'on cherche à trouver dans l'image comme le montre la figure 2.7. Les *patchs* sont passés sur toute l'image et s'activent si l'objet est trouvé. Par exemple, un *patch* générique d'une bouche effectuant un sourire peut être passé sur l'image. On met la valeur d'activation en corrélation avec les *patchs* des yeux issus également des images de sourire [ZT11]. Le problème majeur autour des *patchs* est de trouver le *patch* universel qui se comporterait bien quelle que soit la forme du visage. Ces approches sont à mettre en relation avec les réseaux de convolutions (*CNN*) présentés dans la section 2.2.2. En effet, on peut observer dans les dernières couches d'un réseau convolutif que l'optimisation produit des filtres représentant chaque partie du visage avec différentes expressions, un peu comme un *patch* mais dont les valeurs sont trouvées par apprentissage et complètement automatiquement.

Classification

La classification est l'étape finale du système de la reconnaissance d'expressions faciales dans laquelle le classifieur catégorise les expressions. La figure 2.5 présente les étapes de classification classique pour la reconnaissance d'expressions faciales. La dernière étape de la classification standard comporte deux phases. La première étape consiste à entraîner le classifieur sur des données avec un modèle d'apprentissage. Les bases de données comportent les labels pour chaque image ou vidéo observée. On parle d'apprentissage supervisé. Lors de l'apprentissage, le classifieur sélectionne les descripteurs les plus pertinents afin de correctement séparer les classes de labels. Lors de la seconde étape, le label d'une nouvelle observation est obtenu en utilisant le classifieur entraîné. Pour cela, les descripteurs extraits sont donnés au classifieur qui calcule en retour le label de cette observation. La première étape d'apprentissage peut être plus ou moins longue en fonction de la taille du jeu de données d'apprentissage et aussi du nombre de descripteurs présents. La seconde étape se calcule généralement en temps réel.

2.2.2 Approches à base de réseaux de neurones

Dans les approches qualifiées de "classiques", un humain (un chercheur ou un ingénieur) cherche à produire les descripteurs les plus pertinents possibles en utilisant ses connaissances a priori du problème. Les descripteurs sont issus du traitement d'images et de la reconnaissance des formes traditionnelles, puis sont fournis à un classificateur. Les approches basées sur des réseaux de neurones (*NNB Neural Network Based*) ont révolu-

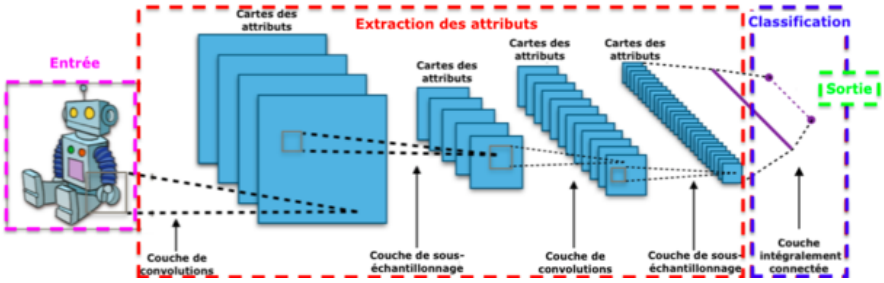


FIGURE 2.8 : L'apprentissage d'un réseau convolutif consiste à optimiser un ensemble de filtres, ainsi que les poids des couches de neurones complètement connectées réalisant la classification finale. Source : *wikipedia*

tionné ce processus en incluant dans un unique processus d'optimisation l'extraction des descripteurs et la classification [LKF10]. Les approches à base d'apprentissage profond ont fait une percée considérable au cours des cinq dernières années, notamment justifiée par l'émergence des approches convolutionnelles *CNN*. Dans les approches *NNB* à base de convolutions, une succession de filtres de convolution sont trouvés par optimisation (apprentissage) à partir des données. Un schéma de réseau convolutif est donné sur la figure 2.8.

Les réseaux convolutionnels

Un réseau convolutif (*CNN*) se compose de trois types de couches : les couches de convolution, les couches de mise en commun et les couches entièrement connectées. La couche convolutive *CNN* se compose d'un ensemble de filtres appris qui convoluent une image et produisent des cartes de caractéristiques. L'opération de convolution a trois avantages : la connectivité locale apprend la corrélation entre les pixels voisins ; le partage des poids dans la même carte de caractéristiques réduit considérablement le nombre de paramètres ; et les filtres sont invariants à l'emplacement de l'objet. Les couches de regroupement sont similaires à la première couche convolutive, mais prennent en entrée des images de caractéristiques. Ces couches sont utilisées pour réduire la taille spatiale des cartes de caractéristiques. La figure 2.9 montre que les derniers filtres s'activent pour des motifs très similaires aux motifs présents sur les images d'entraînement. Les couches entièrement connectées sont à la fin du réseau et servent à classifier, c'est-à-dire à produire le label final.

Ces méthodes *NNB* ont fait progresser les taux de bonnes reconnaissances d'expressions du visage au-dessus de 90% sur les bases de données très contrôlées. Le défi a alors été de travailler sur des problèmes plus difficiles, allant avec la création de nouvelles bases de données dites "sauvages" (*"in the wild"*) [Dha+09 ; Dha+11 ; Dha+12]. Ces bases de données contiennent des images ou des vidéos capturées avec très peu d'éléments contrôlés tels qu'explicités dans la section 2.2.3.

Transfert d'apprentissage

La généralisation des réseaux de neurones profonds a permis le développement de l'apprentissage par transfert. L'apprentissage par transfert est une méthode d'apprentissage

[PY09 ; WKW16] où un modèle développé pour une tâche est réutilisé comme point de départ pour un modèle sur une seconde tâche. Il s'agit d'une approche populaire dans l'apprentissage profond où les modèles pré-entraînés sont utilisés pour initier un modèle visant des tâches plus spécifiques. Cela permet de factoriser les vastes ressources en temps de calcul nécessaires pour développer certains modèles de réseaux neuronaux. L'apprentissage par transfert permet aussi de travailler sur des tâches nouvelles ayant moins de données disponibles [Ng+15]. Ceci ne fonctionne que si les caractéristiques apprises lors de la première tâche sont assez générales ou similaires à celles nécessaires pour réussir la seconde tâche. Dans nos travaux, l'apprentissage par transfert est utilisé afin de proposer une méthode de reconnaissance automatique d'expressions faciales sur une nouvelle base de données d'expressions spontanées d'enfants en partant d'un réseau pré-entraîné sur des images plus générales.

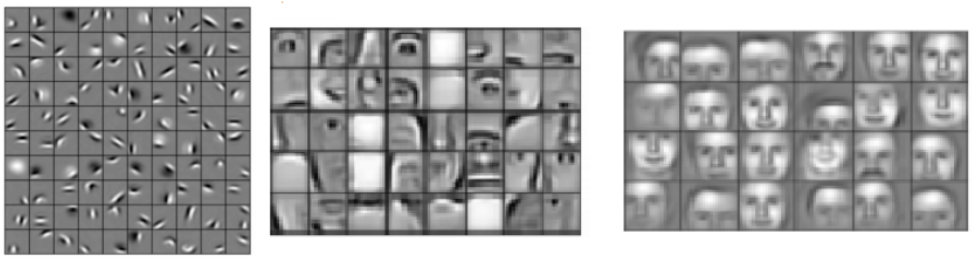


Figure 2.9 : Dans un réseau convolutif, les différents filtres sont trouvés par optimisation (apprentissage). Dans le cas d'un réseau entraîné sur des visages, les premiers filtres sont génériques et détectent des caractéristiques locales. Les derniers filtres cherchent des parties de visages complètes. Source : [Can+22]

Réseaux plus complexes

Les méthodes basées sur les *CNN* ne reflètent pas les variations temporelles des composantes faciales. Des approches hybrides récentes combinent les *CNN* pour les caractéristiques spatiales des images individuelles à des réseaux de neurones récurrents (*Recurrent Neural Network - RNN*), qui sont particulièrement adaptés au traitement de séquences de données [Zha+17]. Puis, les réseaux *LSTM* (*Long Short Term Memory*) comportant une mémoire à court et long terme pour garder les caractéristiques temporelles des images consécutives sont développés. Les *LSTM* sont un type spécial de *RNN* capables d'apprendre les dépendances temporelles.

Les Auto-Encodeurs (*AE*) sont une forme de réseaux de neurones utilisée pour l'apprentissage non supervisé. Les réseaux comportent un encodeur et un décodeur. L'encodeur produit un code latent. Le décodeur est entraîné pour reconstruire les données d'entrée à partir du code latent. L'objectif est de représenter les données d'entrée en une représentation plus efficace. Tout code de l'espace latent produit une donnée de sortie plausible. Un autre type de réseaux capables de représenter efficacement des données dans un espace latent sont les *GAN*(*Generative Adversarial Network*). Les *GAN*reposent sur la mise en compétition de deux réseaux : un générateur et un discriminateur. Le générateur, comme le décodeur d'un *AE*, produit une donnée de sortie plausible. Les *AE* et les *GAN* peuvent

être utilisés pour la reconnaissance d'expressions de visages en utilisant ce code latent pour aider la classification [Zha+18a ; Cai+21].

Les réseaux d'attentions prennent une place de plus en plus importante dans la communauté de vision par ordinateur. Ces réseaux ont été introduits par la communauté de traitement automatique des langues, car dans un texte l'ordre des mots peut-être variables et à de l'importance. Pour les expressions, les réseaux d'attentions offrent la capacité de traiter les vidéos d'expressions [Men+19 ; Gan+20] en se concentrant sur certaines parties de la séquence afin de les traiter avec plus d'attention. Cela est réalisé en ajoutant une couche dite d'"attention" au modèle, qui calcule un poids pour chaque élément de la séquence en fonction de sa pertinence pour la tâche en cours. Ces poids sont utilisés pour pondérer les différents éléments de la séquence lors de la prise de décision finale. Cet état de l'art récent [Can+22] donne un bon aperçu des dernières avancées.

2.2.3 Bases de données d'expressions faciales

Le pipeline global pour la reconnaissance d'expressions faciales a été présenté précédemment en insistant sur les différentes méthodes d'extraction de descripteurs et de classifications [CSB11]. En complément, l'étude des bases de données existantes [Web+18] donne une bonne indication des difficultés et verrous qui sont ciblés au fil du temps. Les caractéristiques qu'il est important de considérer quand on travaille avec ces bases sont les suivantes.

- Individualité des sujets. La forme du visage, la texture de la peau (poils par exemple), lunettes de vue ou de soleil, le sexe, l'origine ethnique et l'âge des sujets sont des paramètres sur lesquels on peut jouer pour créer une base de données variée ou non.
- Expression actée ou expression spontanée. La majorité des bases de données est réalisée en demandant à des personnes de jouer une série d'expressions. Ces expressions dirigées peuvent différer dans leurs caractéristiques, leur dynamique temporelle et leur spontanéité [SF18].
- Images ou vidéos en quelle quantité. Les plus grandes bases de données actuelles comportent au maximum un million d'images et sont issues de requêtes *Internet*. Cela donne des images avec peu de contrôles qui sont parfaites pour relever les défis actuels visant à rendre la reconnaissance d'expressions réellement utilisable, mais les vidéos donnent plus d'informations au prix d'une capture et d'un stockage plus compliqué.
- Résolution de la séquence d'images. Il est nécessaire de faire varier le format des images. Pour cela, certaines bases de données contiennent des images à haute résolution et à basse résolution dans le but de se rapprocher du monde réel.
- Orientation tête/visage. L'orientation du visage par rapport à la caméra influence énormément la performance de différents algorithmes de reconnaissances d'expressions faciales.
- Complexité de l'arrière-plan. La séquence d'images enregistrée avec un arrière-plan complexe rend la tâche de la reconnaissance automatique des expressions faciales encore plus difficile, car l'arrière-plan complexe influence la précision de la détection automatique des visages, le suivi des caractéristiques et la reconnaissance des expressions. La plupart des bases de données disponibles ont un fond neutre ou très persistant.

— Variation d'éclairage. Il est souhaitable que les algorithmes de reconnaissance automatique des expressions soient invariables en fonction des conditions d'éclairage. Très peu de bases de données accessibles au public enregistrent les stimuli d'éclairage variable.
— Type d'expressions. Les expressions vont des six expressions basiques en s'élargissant vers des expressions plus subtiles, notamment en s'intéressant même maintenant aux micro-expressions, difficilement identifiables par un humain non expert.

De nombreux travaux sur la reconnaissance d'expressions faciales utilisent les bases de données suivantes : *Cohn-Kanade (CK) [KCY00], Extended Cohn - Kanade (CK+) [Luc+10], Japanese Female Facial Expressions (JAFFE) [KLG97], MMI [VP10], Multimedia Understanding Group (MUG) [APD10], Yale [GBK01], AR face database [MAR98]*, etc. Les bases de données spécifiques aux enfants sont : *NIMH* [Egg+11], Dartmouth [DGD13], CAFE [LT15]. Des bases de données dites *"in the wild"* sont : *EmotioNet* [FSM16], *Aff-Wild* [Zaf+17], *AffectNet* [MHM17] et *FERV39k* [Wan+22b]. Une liste exhaustive des bases de données est disponible dans l'état de l'art de Li *et al.* [LD22].

Figure 2.10 : La reconnaissance d'expressions faciales cherche à traiter des images prises dans un environnement non contrôlé. Source : [Zaf+17]

2.2.4 Bilan de l'existant en reconnaissance d'expressions faciales

Pour conclure à propos de la reconnaissance d'expressions faciales, les techniques et les bases de données de tests ont beaucoup évoluées. L'un étant lié à l'autre, lorsqu'un verrou semble atteint avec des taux de reconnaissances commençant à être supérieurs à 90%, de nouvelles bases de données sont produites pour aborder des verrous plus complexes. L'objectif final est de pouvoir utiliser les approches dans des applications réelles où aucun contrôle ne sera possible : éclairage variable, nombreuses occultations, tout type de personne, résolution des images faibles, expressions fines, spontanées et subtiles, etc. On peut observer qu'à l'origine les méthodes travaillaient sur des bases de données très contrôlées alors que maintenant les verrous perfectionnés se trouvent dans les spécificités. À travers différents concours, les méthodes se sont améliorées, notamment grâce aux capacités de généralisation dont disposent les réseaux de neurones. Depuis quelques années, des bases dites "sauvages" (voir la figure 2.10) proposent des configurations plus difficiles : visages partiellement cachés, particularités de personne (enfants, paralysie, etc.), faible résolution des images, expressions moins courantes comme la douleur ou les micro-expressions, etc. Dans nos travaux présentés dans les sections 3.1 et 3.2 des cas spécifiques et difficiles sont abordés avec des travaux sur la détection d'expressions à partir d'images

à faible résolution, la détection de la douleur, ainsi que des travaux spécifiques aux visages d'enfants. Les résultats récents [Zha+21] laissent penser que de nombreuses applications peuvent maintenant commencer à utiliser la technologie de reconnaissance d'expressions faciales avec de bons résultats pour des applications courantes. Cependant, de nombreux cas particuliers restent à traiter comme les micro-expressions [Ben+21 ; Aou+21] ou la compréhension des biais présents dans les bases de données [LD20].

2.3 Analyse et reconnaissance d'expressions corporelles

En complément de la parole et des expressions du visage, le domaine de la psychologie a montré que les expressions corporelles sont aussi puissantes que les expressions faciales pour exprimer des émotions [Meh68]. Même dans le cas de situations comportant de nombreux stimuli, la posture permet d'améliorer la reconnaissance des expressions [Bui+14]. Les Sciences Humaines au sens large ont été les premières à travailler sur les relations entre les mouvements du corps et les expressions. De nombreux chercheurs en sciences humaines cherchent à comprendre quels facteurs induisent quels sentiments afin de mieux comprendre les interactions humaines [Mon18]. Les arts visuels ont également tenté de formaliser la relation entre le geste et l'émotion pour améliorer la gestuelle des personnages de dessins animés [TJT95], ou la chorégraphie des spectacles vivants, [LU71 ; VCH94]. En informatique, l'expression corporelle est encore un domaine relativement peu exploré, même si des travaux existent depuis longtemps, le nombre croissant de publications ces dernières années laisse penser que de nombreuses investigations sont encore possibles.

Avec la prolifération et la popularité croissantes des dispositifs de capture de mouvements de squelette humain, comme la Kinect [Han+13], la capture basée sur des accéléromètres [AWS19], ou plus récemment la capture monoculaire [Cao+19 ; Des+21 ; Con20 ; Lug+19], etc., les chercheurs aussi bien en psychologie qu'en vision par ordinateur se sont tournés petit à petit vers l'analyse des mouvements représentés par un squelette en mouvement. Alors que pour le visage, la vidéo offre deux types d'information à analyser : les mouvements des arêtes (nez, lèvre, sourcils, etc.) et la texture, l'analyse des mouvements du corps peut se contenter des mouvements du squelette qui est plus léger que le traitement des vidéos. La technologie de capture de mouvements consiste à trouver le squelette dans les vidéos. Le domaine de la vision par ordinateur s'est d'abord intéressé au problème de la reconnaissance d'actions basée sur l'analyse du squelette [WRB11 ; Bar+14 ; Che+19 ; KF22 ; Bou+21]. Et plus récemment, la communauté s'est intéressée à la reconnaissance des expressions à partir des mouvements du corps [Nor+18].

Cette section détaille d'abord les travaux autour de l'analyse des mouvements expressifs en sciences humaines, puis présente les travaux en reconnaissance d'expressions en vision par ordinateur. Les travaux du domaine de synthèse de mouvements travaillant sur les expressions et styles en informatique graphique seront présentés dans la section suivante.

Figure 2.11 : Des exemples de différentes postures associées aux six émotions de bases. Source : [SVD08]

2.3.1 Analyse des expressions corporelles

De nombreux travaux en psychologie et en sciences humaines cherchent à comprendre et à catégoriser les relations entre la posture, le mouvement et les expressions. L'objectif est d'interpréter quels phénomènes induisent quels sentiments afin de mieux comprendre le fonctionnement humain.

Posture et mouvements

Vaina *et al.* [Vai+90] et par Giese et Poggio [GP03] ont montré qu'il existe deux voies séparées dans le cerveau pour reconnaître des informations de postures et de mouvements. La première voie se concentre sur l'information de forme en cherchant à analyser la posture du corps, tandis que la seconde voie va étudier les informations de mouvements afin de reconnaître l'action ou l'expression. D'autres études [McL96 ; HH06 ; PWD06] ont confirmé cette hypothèse en observant que les informations de forme donc de posture sont cruciales pour la reconnaissance de mouvements corporels. De même, Atkinson *et al.* [ATD07 ; OG07b] montrent que l'information seule de mouvement permet également de reconnaître l'expression corporelle d'un mouvement particulier. En effet, en analysant un stimulus lumineux posé sur une articulation, par exemple sur la main lors d'un mouvement de frapper à une porte, il est possible de discriminer différentes expressions. Cette étude confirme que considérer uniquement le squelette est une hypothèse raisonnable pour effectuer la reconnaissance d'expressions corporelles.

Le modèle de Laban

Le modèle de Laban a été proposé par Rudolf Laban [LU71 ; VCH94], un chorégraphe hongrois, dans le but de caractériser l'expressivité de mouvements de danse. Ses travaux sont utilisés dans de nombreux domaines de recherche tels que la psychologie, la robotique,

l'animation par ordinateur ou l'analyse du mouvement. Une partie importante de son travail est consacrée à la recherche de facteurs permettant de communiquer et de transcrire les qualités expressives d'un mouvement. En effet, il a développé une analyse de mouvement du nom de *Laban Movement Analysis* (*LMA*). Cette analyse se base sur 4 grands axes : le flux (flow), l'espace (space), le poids (weight) et le temps (time). La figure 2.12 à gauche représente ces quatre axes. Dans son étude, Laban définit différents plans du mouvement : le plan de la table (horizontal), de la porte (vertical) et de la roue (sagittal). Il construit une sphère du mouvement, la kinésphère, qui désigne l'espace accessible directement aux membres d'une personne comme illustré sur la figure 2.12 à droite. Elle s'étend tout autour d'elle, jusqu'à l'extrémité de ses doigts et pieds tendus dans toutes les directions. Le modèle de Laban a été étendu et se base maintenant sur six catégories distinctes.

— Le corps. Qu'est-ce qui bouge, et comment ? Quel est le mouvement produit ?
— L'espace. Où va le mouvement ? Dans quel espace s'inscrit-il ?
— L'effort. Comment le mouvement est-il exécuté ? Avec quelles qualités d'énergie ?
— La forme. Quels sont les différents chemins empruntés par le mouvement ?
— Le phrasé ou le rythme. Dans quels laps de temps et suivant quel rythme s'effectue le mouvement ?
— L'inter-relation. Comment l'individu en mouvement est-il en relation avec son entourage ?

Ce modèle est devenu très populaire en vision par ordinateur et en synthèse d'animations pour la reconnaissance d'actions, la reconnaissance d'expressions corporelles et la génération de mouvements expressifs.

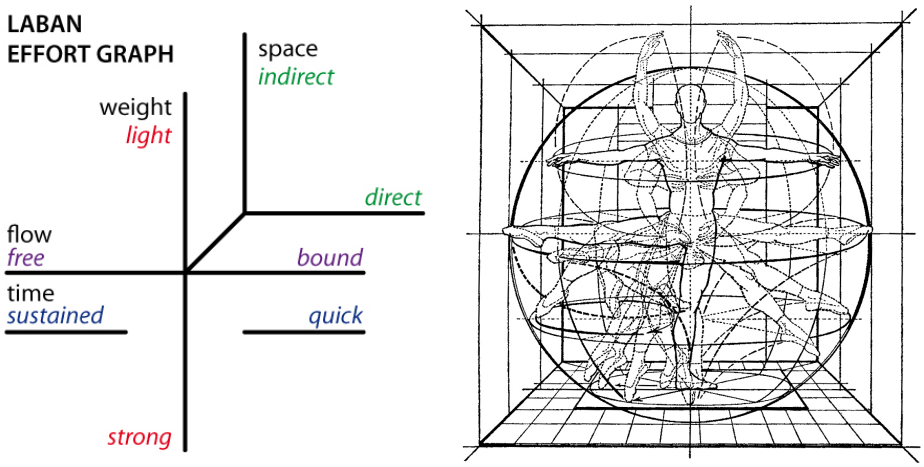


Figure 2.12 : Gauche : diagramme de l'effort selon Laban avec le poids (léger ou fort), l'espace (indirect ou direct), le temps (soutenu ou urgent) et le flux (libre ou contenu). Droite : la kinésphère représente la globalité des endroits de l'espace que l'on peut atteindre lorsque l'on se tient sur un pied.

Définitions de caractéristiques

Orthogonalement à la distinction entre posture et mouvement, de nombreux travaux cherchent à caractériser la posture ou le mouvement pour chaque expression. James

[Jam32a] a découvert que l'ouverture du corps est un critère important lorsqu'un humain cherche à reconnaître une expression corporelle, tout comme l'inclinaison du corps et la position de la tête (penchée en avant, en arrière ou tournée). D'autres travaux [Meh68 ; HR83] ont confirmé l'importance de l'inclinaison du corps afin de reconnaître des expressions corporelles. En utilisant des mouvements de danse, Aronoff *et al.* [AWH92] ont montré que des postures dites arrondies vont être plus chaleureuses tandis que des postures avec des configurations angulaires et diagonales vont signifier un comportement menaçant. De Meijer [Mei89] a été l'un des premiers à construire une étude comportant plusieurs caractéristiques et plusieurs expressions. Il voulait déterminer si des mouvements corporels spécifiques de danse étaient indicatifs d'émotions spécifiques et quelles caractéristiques du mouvement étaient à l'origine de ces attributions. Wallbott [Wal98] a étendu ces travaux en traitant des expressions non jouées et l'aspect interculturel.

Au cours des années 2000, grâce à la possibilité d'automatiser l'analyse, les chercheurs du domaine de la psychologie ont pu améliorer leur inventaire de caractéristiques (appelés descripteurs en vision) qui permettent aux humains de distinguer les émotions spécifiques. Par exemple, en utilisant une approche fondée sur la théorie de l'information, De Silva et Bianchi-Berthouze [DB04 ; KDB06 ; KB07] ont proposé un jeu de descripteurs permettant la discrimination de quatre émotions. Vingt-quatre descripteurs sont proposés pour décrire la position des articulations du haut du corps et l'orientation des épaules, de la tête et des pieds. L'analyse statistique a montré que peu de dimensions étaient nécessaires pour expliquer la variabilité de la configuration des descripteurs. Gross *et al.* [GCF10] ont analysé le mouvement de "frapper à une porte". Leur étude a montré que la vitesse et l'accélération de la main permettent de reconnaître l'expression corporelle d'une personne pour les expressions à forte activation, c'est-à-dire la joie, la colère et la fierté. Ils ont quantifié les valeurs des différentes caractéristiques sur différentes expressions corporelles. Les résultats ont été positifs, ce qui signifie qu'une comparaison quantitative des expressions est possible. Par exemple, ils ont constaté que le bras est levé au moins 17 degrés plus haut pour les mouvements de colère que pour les autres expressions. Glowinski *et al.* [Cam+11] ont montré que l'utilisation d'un ensemble réduit de caractéristiques, du haut du corps seulement, est suffisante à un humain pour reconnaître un grand nombre de comportements affectifs [BMS12].

Plus récemment, la possibilité de montrer des mouvements virtuels a permis de créer des expériences nouvelles. Par exemple, Normoyle *et al.* [Nor+13] cachent les visages pour que l'observateur se focalise sur les mouvements corporels (voir la figure 2.13). Giraud *et al.* [Gir+15] ont étudié les inférences faites par des observateurs concernant les émotions spontanées et les traits de personnalité des entraîneurs humains à partir des mouvements cinématiques qu'ils produisent lors de l'exécution d'une séance de fitness. Les humains testés visualisaient des mouvements d'un avatar virtuel. Les perceptions des observateurs étaient partiellement exactes pour les dimensions émotionnelles et ne l'étaient pas pour les traits de personnalité. Les résultats identifient des caractéristiques de mouvement qui contribuent à la compréhension de la perception sociale par le mouvement. Plus récemment, la définition des descripteurs de mouvements s'est souvent associée aux approches de reconnaissance automatique d'expressions corporelles décrites dans la section 2.3.2.

Figure 2.13 : L'animation d'avatar permet de nouvelles expériences pour comprendre comment un humain reconnaît une expression. Ici, l'expression du visage ne perturbe pas la perception des mouvements du corps. Source : [Nor+13].

Corpus non joué

Il est à noter que l'étude de Kleinsmith *et al.* [KDB06] aborde la question de l'obtention de postures affectives non actées. Ils ont collecté des données de mouvements de personnes jouant à des jeux de sport sur la *Nintendo Wii* et ont utilisé des postures statiques à partir des données lorsqu'un point a été gagné ou perdu dans le jeu. Ensuite, chaque posture a été associée à un vecteur contenant une description de la posture au niveau le plus bas. Une analyse statistique des descripteurs a montré que les descripteurs les plus importants étaient principalement les bras et le haut du corps. Bien qu'il y ait eu une discrimination significative entre les quatre émotions distinctes, une plus grande discrimination a été obtenue entre les états affectifs plus "actifs" (frustrés et triomphants) et les états moins "actifs" (concentrés et défaite). Il existe peu de bases de données d'expressions corporelles spontanées hormis celle-ci. Pour certaines applications, notamment en synthèse, cela n'est pas forcément gênant, car souvent l'animateur se satisfait du côté exagéré des expressions actées.

Bilan des travaux en analyse des mouvements expressifs

En général, les différentes études du domaine de la psychologie montrent que l'analyse de la posture peut permettre une discrimination par un humain de différents états affectifs (voir la figure 2.11). Cependant, cela est loin de signifier qu'il existe une relation unique entre une émotion discrète et une expression corporelle. De même, la plupart des études sur les expressions corporelles semblent également présenter des schémas assez discriminatoires lorsqu'on considère des combinaisons de descripteurs plutôt que des descripteurs individuels. Par exemple, Wallbott [Wal98] montre que dans certains cas, comme pour le dégoût et la fierté, les bras vont être croisés, mais la caractéristique discriminante semble être la position de la tête (penchée vers l'avant pour le dégoût et vers l'arrière pour la fierté). Toutes ces études renforcent l'idée qu'il existe des facteurs contextuels qui peuvent affecter la façon dont un état émotionnel est exprimé. Tout cela semble en accord avec Russell [Rus03], qui affirme que les expressions prototypiques sont en fait assez rares. Les descripteurs qui forment l'expression d'une émotion ne sont pas fixes et chaque réaction émotionnelle n'est pas unique. Un travail de fond est donc nécessaire afin de produire des approches automatiques de reconnaissance présentées dans la sous-section suivante.

Une autre question importante qui ressort immédiatement de cet état de l'art est le manque de vocabulaire commun utilisé par le domaine de la psychologie pour décrire les des-

cripteurs. Les descriptions semblent souvent fondées sur des évaluations subjectives et qualitatives et sont donc difficiles à comparer d'une étude à l'autre. De plus, les descripteurs de haut niveau sont très dépendant du contexte de l'action réalisée et difficile à comparer sans décomposer et interpréter les termes. Plusieurs de nos travaux vont s'inscrire directement dans la continuité de ces constats.

2.3.2 Reconnaissance automatique d'expressions corporelles

Comme présenté dans la section 2.3.1 de ce chapitre, les chercheurs en psychologie et sciences humaines au sens large ont été les premiers à s'intéresser aux expressions corporelles portées par la posture et le mouvement du corps. Très vite, ils se sont concentrés sur l'analyse des mouvements représentés par un squelette comme illustré par la figure 2.14, car elle est plus abordable en taille de données que l'analyse vidéo [She+19], parce que leurs expériences ont montré que visuellement un observateur retrouvait l'expression avec juste un squelette et parce que les systèmes de captures de mouvements se sont démocratisés. Leurs travaux sur la définition de caractéristiques sont le point de départ des techniques de reconnaissance automatique d'expressions corporelles.

Cette section les présente, en commençant très rapidement par les travaux en reconnaissance d'actions qui traitent également un squelette animé [Bar+12]. Notre taxonomie essaie de regrouper les approches par différentes caractéristiques. Bien que ce domaine soit relativement récent, il est très difficile d'être exhaustif, notamment à cause de la variété de domaines traitant du sujet. Il existe plusieurs états de l'art regroupant les travaux spécifiques aux mouvements du corps avec l'article de Kleinsmith *et al.* [KB13], puis celui de Zacharatos *et al.* en 2014 [ZGC14], de Noroozi *et al.* en 2018 [Nor+18] et de Ribet *et al.* [RWV19]. Il est intéressant aussi de se référer à l'article de Stephens-Fripp *et al.* [Ste+17] spécifique à la démarche, ainsi qu'à celui de Wang *et al.* [Wan+22a] traitant de la fusion des données du corps, du visage, de capteurs physiologiques et de la parole.

Reconnaissance d'actions

La reconnaissance d'actions est une problématique connexe et antérieure à la reconnaissance d'expressions. La communauté de vision par ordinateur a longtemps travaillé directement sur les vidéos pour reconnaître une action [KF22]. Avec les récentes avancées en capture de mouvements monoculaires [Cao+19 ; Con20 ; Lug+19], la problématique de reconnaître une action se réalise souvent à partir des mouvements d'un squelette [PL16 ; Che+19]. Cette problématique partage avec la reconnaissance d'expressions le défi de devoir traiter une animation, avec de nombreux degrés de liberté. Néanmoins, les approches de reconnaissance d'actions ne peuvent pas être utilisées directement pour reconnaître les expressions corporelles. Une expression peut être considérée comme un enrichissement d'une action, d'un geste. Dans la reconnaissance des expressions corporelles, les actions et les expressions se chevauchent, se mélangent. La reconnaissance des expressions est un problème orthogonal à la reconnaissance d'actions, puisque l'expression modifie le geste de manière subtile et à peine perceptible. Afin de simplifier un peu le problème, de nombreux travaux n'ont traité que des mouvements spécifiques à une unique action. Même si séparer l'action et l'expression ne semble pas trivial, c'est intuitivement un défi intéressant à relever et nos travaux de la section 4.2 proposent de s'y atteler.

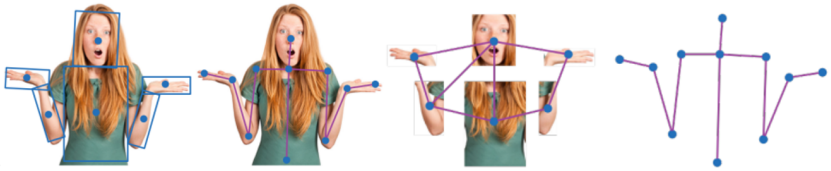


Figure 2.14 : Les deux façons les plus courantes de modéliser le corps humain pour reconnaître son expression sont soit à partir de la vidéo, soit à partir d'un modèle cinématique appelé squelette. Depuis l'image, les différentes parties du corps sont détectées pour obtenir le squelette. Contrairement au visage, la texture du corps donne peu d'information sur l'expression, un humain reconnaît l'expression simplement à partir du squelette. Source : [Nor+18].

Mouvements spécifiques

Pour essayer de simplifier le problème du mélange de l'action et de l'expression, de nombreux travaux se sont concentrés sur l'étude d'un mouvement spécifique pour pouvoir faire certaines simplifications ou trouver plus facilement des régularités. Les mouvements spécifiques étudiés sont par exemple la marche [Roe+09a ; KKB10 ; Bar+13 ; OG07b ; Bha+20 ; Ran+22 ; Lia+22], l'action de frapper [BR07 ; GCF10], les orateurs [Vol+14], les mouvements de la main [KM09] ou les performances artistiques [Sen+16]. Parmi les actions spécifiques, la reconnaissance d'expressions dans des mouvements de danse est souvent abordée [Cam+03 ; Par+04 ; CTV02 ; Sen+16 ; Ari+15].

Ces approches proposent des descripteurs en s'inspirant des travaux d'analyse de mouvements décrits précédemment. Des descripteurs de bas niveau se réfèrent aux articulations : angle de rotation, position 3D, vitesse, accélération, distance entre articulations, etc. Les descripteurs de haut niveau comme ceux issus des modèles de Laban présenté dans la section 2.3.1 sont souvent peu élaborés dans ces approches. Ces méthodes sont un premier pas, mais sont souvent trop spécifiques à une action. Elles échouent lorsque le défi est de reconnaître des expressions corporelles de différents scénarii.

Mouvements hétérogènes et descripteurs experts

Pour traiter des mouvements hétérogènes comportant tous types d'actions amplifiées par une expression, les descripteurs doivent être plus complexes. Dans ce document, nous utiliserons souvent le terme de descripteurs experts, car ils sont élaborés en se basant sur les connaissances des experts en mouvements expressifs (scientifiques, chorégraphes, animateurs, etc.). Ces descripteurs qualifiés d'experts peuvent provenir de multiples familles : cinématiques, géométriques, fréquentiels, statistiques, etc. Par exemple, Dael *et al.* [DGS13] montrent que la quantité de mouvement, la vitesse, la force, la fluidité, la taille et la hauteur/position verticale sont des caractéristiques importantes de la posture qui reflètent l'émotion. Barliya *et al.* [Bar+13] indiquent que le bonheur et la colère sont principalement exprimés par une augmentation de la vitesse de mouvement, des oscillations du bras et de la cadence. Inspirés par la psychologie et les théories humanistes, Piana *et al.* [Pia+16] extraient l'indice de contraction, la fluidité et l'impulsivité des mouvements de la posture. Fourati *et al.* [FP15 ; FPD19] étudient spécifiquement la contribution de différents types de caractéristiques temporelles à partir de leur large base de données de mouvements

expressifs [FP14]. Larboulette *et al.* [LG15] proposent quant à eux une revue des méthodes de calcul des descripteurs experts.

Nos premiers et derniers travaux dans ce domaine de la reconnaissance d'expressions corporelles présentés dans le chapitre 4 s'inscrivent dans cette famille de méthodes basées sur des descripteurs experts. Dans ces travaux, tous les descripteurs existants sont regroupés dans un unique formalisme, puis cette voie est poussée plus loin avec une étude qualitative associant des valeurs numériques aux descripteurs. Ces descripteurs ont l'énorme avantage par rapport aux approches plus récentes basées sur l'apprentissage automatique, d'être compréhensibles facilement, pour pouvoir les utiliser également en synthèse d'animations.

Descripteurs inspirés du modèle de Laban

Parmi les descripteurs cités précédemment, le formalisme du modèle de Laban conçu par le chorégraphe pour la danse a une place particulière, car il est très souvent transposé aux gestes non dansés avec pour objectif de traiter des actions hétérogènes. Par exemple, Truong *et al.* [TBZ16] ciblent d'abord la reconnaissance d'actions. Cependant, ils l'appliquent aussi à la reconnaissance d'expressions sur leur base de données de mouvements spécifiques. Les résultats obtenus sont très prometteurs, mais les mouvements expressifs évalués sont assez limités en termes de diversité d'actions. Dewan *et al.* [DAS17] étendent le modèle en le combinant avec une fenêtre temporelle afin de segmenter le mouvement en entrée. Le modèle de Laban est la brique de base de nombreux travaux du domaine [Zac+13b ; AMD19]. Son influence pour les gestes du corps est comparable au modèle *FACS* d'Ekman pour les visages présenté dans la section 2.1.3.

Approches récentes et apprentissage profond

Dans les domaines de la reconnaissance d'actions [KF22] et de la reconnaissance d'expressions faciales, les avancées très récentes se sont principalement basées sur des approches d'apprentissage profond [LD22] (voir également la section 2.2.2). Néanmoins, pour la reconnaissance d'expressions corporelles, il y a globalement un manque de grands ensembles de données, qui sont nécessaires pour l'apprentissage profond. Même très récemment, les recherches s'appuient toujours sur des descripteurs experts [Nor+18]. Il existe un groupe de méthodes utilisant l'analyse statistique et automatique, qui n'exploitent pas les réseaux de neurones pour trouver des régularités. Par exemple, ces deux approches [Kac+18 ; Dao+17] utilisent les matrices de covariance pour encoder les corrélations spatiales entre les articulations pendant les mouvements gestuels humains et exploitent les propriétés géométriques des matrices de covariance. Nos travaux de la section 4.2 se proposent de séparer l'action de l'expression pour ne classifier que l'essentiel. Ces approches récentes sont plus élaborées que les descripteurs experts, mais n'utilisent pas de réseaux de neurones qui semblent être maintenant la brique de base incontournable.

Les approches basées sur les réseaux ne font leur apparition que depuis quelques années. Souvent, un lien est fait avec les images. Par exemple, Santhoshkumar *et al.* [SG19] utilisent un réseau de neurones convolutif (*CNN*) sur les vidéos des personnes exprimant une émotion, sans passer par une animation à base de squelette. Cependant, les approches effectuent souvent une première passe pour retrouver le squelette qui est alors analysé.

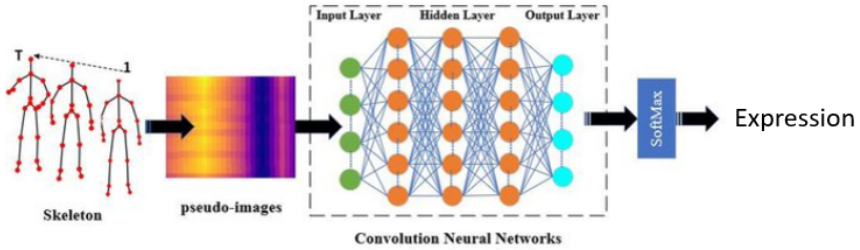


Figure 2.15 : Les animations basées sur squelette sont souvent transformées en une image de données afin de pouvoir être utilisées avec un réseau convolutif (*CNN*).

Beyan *et al.* [Bey+21] proposent une fenêtre glissante pour segmenter les données de pose et les transformer en une image de données afin de pouvoir utiliser les réseaux convolutifs 2D classiques comme illustrés sur la figure 2.15. Ly *et al.* [Ly+18 ; Liu+21] extraient les caractéristiques temporelles des gestes liées à l'émotion en utilisant l'analyse convolutive temporelle : *LSTM*. Karumuri *et al.* [Kar+19] proposent des convolutions 1D sur chaque canal d'une animation pour classifier quatre émotions. Shi *et al.* [Shi+22] introduisent un réseau de convolution multi-échelles à base de graphe. Un squelette humain peut être assimilé à un arbre, l'utilisation des réseaux de neurones avec une topologie de graphe semble donc bien appropriée [Wu+20]. Il existe également des travaux sur de nouveaux types de réseaux. Par exemple Huang *et al.* [HV17] proposent un nouveau réseau riemannien pour extraire l'information spatiale du geste. Parmi les applications, ils montrent un exemple de reconnaissance de geste 3D expressif. Les réseaux d'*attentions* (*transformer* [Vas+17]) qui agitent beaucoup la communauté de vision actuellement, semblent pouvoir aider à résoudre de nombreux problèmes de reconnaissances. Qin *et al.* [Qin+22] *et al.* les utilisent en reconnaissance d'actions. Il est probable que ces réseaux améliorent également la reconnaissance d'expressions corporelles. Dans l'ensemble, les méthodes à base d'apprentissage profond améliorent les taux de classifications correctes par rapport à celles basées sur des descripteurs experts, mais au détriment de l'explicabilité. Des efforts de la communauté sont à réaliser dans cette direction, ainsi que pour augmenter la quantité des données disponibles.

2.3.3 Données de mouvements corporels expressifs

Nous avons validé nos différentes méthodes de reconnaissance d'expressions corporelles sur différentes bases de données afin de tester la généricité et la robustesse des approches proposées dans ce manuscrit. Les bases de données sont *Emilya* [FP14], *UCLIC* Affective Body Posture and Motion [KDB06], *MPI* Emotional Body Expressions Database for Narrative Scenarios [Vol+14], Biological Motion [MPP06] et la base de données provenant d'animateurs [Xia+15a] et non de capture de mouvements que l'on nommera *SIGGRAPH-DB* dans la suite de ce manuscrit. Leurs caractéristiques sont détaillées dans la thèse de Crenn [Cre19]. Il existe d'autres bases de données, souvent propriétaires [TBZ16 ; Zac+13a], des bases spécifiques à la démarche [Bha+20 ; Ran+20], à la personnalité [Dur+09] et des bases multimodales [Sap+18].

Représentation des expressions. La grande majorité des approches de reconnaissances cherchent à identifier les labels discrets d'expressions liés souvent aux six expressions de base du modèle d'Ekman présenté dans la section 2.1.3. Ces labels sont souvent ceux de la base de données utilisée. Très peu de travaux et de bases traitent d'autres représentations, notamment la représentation continue de type arousal/valence qui aurait un intérêt en synthèse d'animations. En effet, pouvoir passer de manière continue d'une expression à une autre offrirait plus de diversité. Il y a un travail à effectuer autour de ce point, mais qu'il faudrait traiter dès la conception de la base de données en ciblant la représentation arousal/valence des expressions.

Expressions actées et spontanées. Comme pour le visage, peu de bases de données proposent des expressions spontanées. Kleinsmith *et al.* [BK03 ; KB07 ; KBS11] offrent la base *UCLIC* avec des postures non jouées et des états affectifs subtils obtenus à partir de personnes jouant à des jeux vidéo. Filntisis *et al.* [Fil+19] ont développé une approche pour analyser une base de données contenant des expressions actées et spontanées d'enfants participant à un scénario d'interaction avec des robots. La base est très spécifique et les mouvements du corps ont servi principalement de complément au visage et aux paroles.

2.3.4 Bilan de l'existant en reconnaissance d'expressions corporelles

Comparé à la reconnaissance d'expressions faciales, le domaine de la reconnaissance d'expressions corporelles propose beaucoup moins de publications, mais la problématique est vouée à se développer avec les nouvelles techniques de capture d'un mouvement avec une simple caméra en temps réel. Les données disponibles devraient s'accroître. Pour l'instant, les bases de données publiques contiennent encore un nombre assez restreint de mouvements avec peu d'expressions différentes, à part *Emilya*. Trop de chercheurs utilisent encore uniquement leur base de données propriétaire afin d'évaluer leur méthode. La base *Emilya* est probablement la base la plus testée pour comparer les résultats.

Les techniques basées sur l'extraction de descripteurs experts obtiennent de bons résultats. Les approches à base de réseaux proposent des résultats meilleurs, mais sans franchir de palier réellement impressionnant. Notre avis est que les données n'offrent peut-être pas assez de défis ni de diversités pour que les réseaux expriment leur potentiel. De plus, les descripteurs experts gardent l'avantage d'être réellement explicables. Il faut donc explorer les deux familles de possibilités et probablement croiser les idées entre les deux mondes. Les réseaux semblent incontournables, mais on peut essayer de les contraindre à produire des transformations de données explicables. Par exemple, Wu *et al.* [Wu+21] utilisent un auto-encodeur sémantique qui convertit des gestes d'une expression non connue en une représentation de plus bas niveau apprise sur des expressions connues. Ce type d'approche donne une idée d'une direction à suivre : utiliser une série de petits réseaux experts en maintenant une compréhension de chaque résultat intermédiaire. Les scientifiques qui analysent les comportements seraient preneurs de ces informations, et ceci permettrait également d'appliquer certaines idées à la synthèse d'animations.

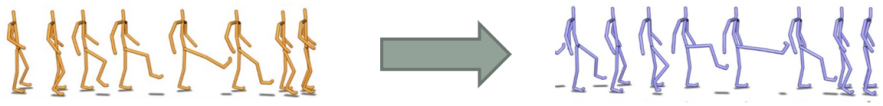


Figure 2.16 : Une animation est représentée par une séquence de poses clés (*keyframing*). Il est possible d'éditer chaque pose pour obtenir une nouvelle animation à droite.

2.4 Animations et styles

La synthèse d'animations de personnage virtuel dispose d'un champ d'application extrêmement vaste : cinéma, jeux vidéo, jeux sérieux, agents virtuels, médecine pour des aspects d'analyse ou de rééducation du geste, etc. Certaines applications comme le cinéma ou le jeu vidéo ont tendance à favoriser le côté plausible et réaliste des animations, alors que d'autres, comme la médecine ou le sport, préfèrent un réalisme physique et précis des gestes. L'état de l'art à suivre se concentre plutôt sur les techniques privilégiant un réalisme plausible à une cohérence physique. Historiquement, les travaux en animation 3D pour le divertissement se sont focalisés sur deux aspects : des outils facilitant la créativité des animateurs et des moteurs permettant de produire en temps réel des animations toujours plus réalistes. La technique de poses clés (*keyframing* en anglais) introduite par l'animation traditionnelle est aussi employée en 3D en manipulant un squelette dont la pose est décrite pour chaque pas de temps, comme le montre la figure 2.16. En effet, il est plus simple de travailler sur le mouvement d'un squelette séparément du maillage 3D. Cet état de l'art ne présente que les techniques d'animation traitant le squelette, de la même manière que dans l'état de l'art en reconnaissance d'expressions corporelles. La déformation du maillage du personnage virtuel associé, ainsi que le calcul du rendu de l'image sont des domaines extérieurs à la problématique de nos travaux.

2.4.1 Édition et production d'animations

Les applications de monde virtuel doivent animer simultanément de nombreux personnages virtuels souvent en temps réel. Les deux grandes familles d'approches sont l'animation procédurale et l'animation basée sur des données. L'animation basée sur des données peut se décomposer en deux familles de problèmes : l'édition d'une animation en passant par l'édition de poses et la génération du mouvement d'un personnage en temps réel dans un monde virtuel.

Animations procédurales

Les approches procédurales se basent sur des assemblages de petites procédures qui résolvent chacune une sous-partie d'un problème et n'utilisent aucune donnée. Elles sont utilisées pour simuler les lois de la physique pour animer les objets en mouvement, comme les particules, les liquides, etc. Pour l'animation de personnage, elles sont utilisées dans les cas où les données manquent ou sont difficiles à obtenir comme pour les animations de la main [Whe+15], pour le langage des signes [NLG20], pour les foules [El-+16], pour les insectes [Guo+14] (voir la figure 2.17) ou pour des créatures avec des morphologies imaginaires [Bha19]. Ces approches peuvent également être liées aux domaines des agents

conversationnels qui cherchent à produire un ensemble de comportements (audio, visage, gestuelle) de personnages virtuels. Certaines approches sont basées sur des machines à états procédurales décrivant des réactions de comportements avec des grammaires [De +04 ; BP04].

La synthèse d'animations procédurale a été beaucoup explorée dans les années 1990 [Mul+99], mais est majoritairement remplacée par des approches à base de capture de mouvements à cause du réalisme des mouvements capturés. Cependant, l'animation procédurale conserve les avantages d'être extrêmement contrôlable, compréhensible par un animateur et très rapide à calculer. Grâce à ces qualités, les approches procédurales sont d'ailleurs maintenant employées comme simplifications de modèles dans les approches cherchant à optimiser les mouvements d'un personnage virtuel plongé dans une simulation physique. Chen *et al.* [Che+22] l'utilisent pour produire le mouvement de papillons plongés dans une simulation physique. Yin *et al.* [YLv07 ; WFH09] l'utilisent dans *SIMBICON* pour contrôler l'animation physique d'un humanoïde. Alvarado *et al.* [Alv+22] l'utilisent pour gérer l'équilibre d'un personnage marchant sur sol meuble.

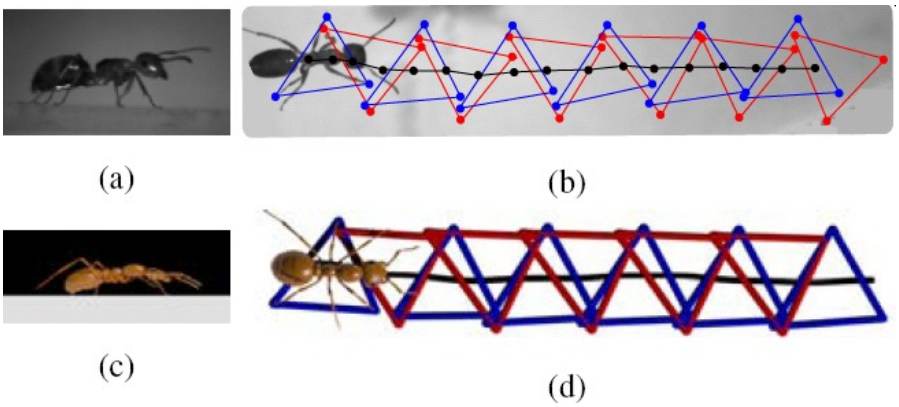


Figure 2.17 : Guo *et al.* [Guo+14] proposent une approche procédurale d'animation de fourmis où les positions des pattes sont une succession de triangles.

Édition d'animations

Grâce à la démocratisation de la capture de mouvements, l'animation basée sur des données représente la grande majorité des techniques d'animation utilisées pour les personnages. Les systèmes d'animation possèdent une énorme base de données de capture de mouvements réalisée à l'aide d'acteurs effectuant tout type de gestes. Le réalisme des animations est l'avantage majeur de ces approches. En revanche, ce qui est capturé n'est souvent pas complètement en phase avec le besoin. L'édition et l'adaptation des animations à un environnement ou un contexte différent sont alors une des problématiques importantes. Les animateurs expriment leur savoir-faire pour produire des animations comportant les effets désirés, mais souvent au prix de processus chronophages. En effet, leurs outils s'appuient sur des outils de bas niveau utilisant la cinématique inverse pour déplacer certaines articulations dans chaque pose clé. Il y a alors un besoin de nouvelles métaphores d'interactions et d'interfaces pour faciliter leur travail. L'autre problématique importante autour des animations basées sur la capture de mouvements est de combiner différentes animations pour pouvoir animer en continu les personnages interagissant avec un environnement.

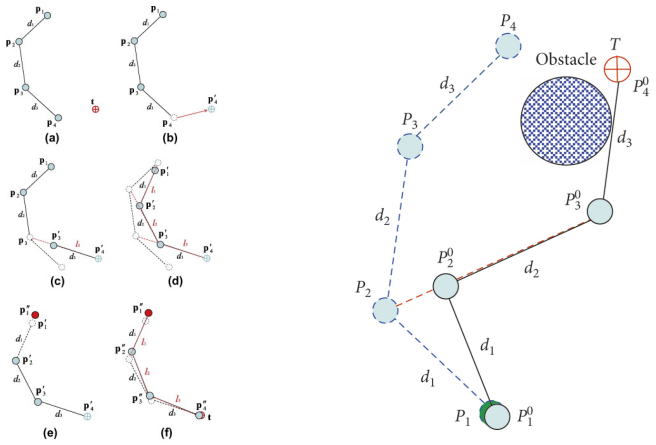


Figure 2.18 : *FABRIK* propose une heuristique efficace pour résoudre le problème de cinématique inverse. Son implémentation simple permet d'ajouter facilement des contraintes comme l'évitement d'obstacle à droite. Source : [AL11].

Cinématique inverse La cinématique inverse (*Inverse Kinematics - IK*) permet de trouver une paramétrisation appropriée d'une chaîne d'articulations en fonction d'un paramétrage cible de certaines articulations, souvent les extrémités. Cette technique peut être utilisée comme une interface pour la manipulation du squelette où les animateurs peuvent faire bouger une articulation sur une position ou une orientation souhaitée tout en maintenant d'autres articulations en place. Le problème de l'*IK* est mal posé, car pour une cible donnée d'une extrémité, il existe une, plusieurs, ou aucune configuration(s) de la chaîne cinématique. Il est possible de répartir les solutions existantes en différentes familles [Ari+18] : analytiques, numériques et à base d'heuristiques.

Les solutions analytiques tentent de trouver une formulation du problème qui donne une solution exacte pour une configuration cible [RR93 ; Kal08]. Elles se basent sur l'inversion d'équation de cinématique directe. Ces solutions fonctionnent bien pour deux ou trois articulations, mais ne sont pas facilement généralisables à n'importe quel squelette à plus de n articulations.

Les solutions numériques formulent une fonction de coût pour le problème et la minimisent par itérations. L'optimisation peut se faire par différentes techniques comme la descente de gradient, l'inversion de la matrice Jacobienne, les moindres carrés, etc. [Bus04]. Le choix de l'approche d'optimisation dépend de la forme de la fonction de coût. Beaucoup de ces approches ont des difficultés à rester stables dans certaines configurations extrêmes. De plus, l'ajout de contraintes supplémentaires comme des bornes d'angles peut être compliqué à intégrer à la fonction de coût.

Les approches à base d'heuristiques s'appuient sur des opérations simples répétées de manière itérative pour atteindre une solution acceptable. Ce processus leur permet d'être extrêmement efficaces sur le plan informatique et facile à comprendre. Leur rapidité en fait la solution actuellement privilégiée des logiciels d'animation. Les heuristiques les plus populaires sont *Cyclic Coordinate Descent (CCD)* [WC91b] et *Forward and Backward Reaching Inverse Kinning* (*FABRIK* [AL11]. L'algorithme *CCD* fonctionne de manière itérative comme illustré sur la figure 2.18 : chaque articulation est tournée de l'angle entre l'effecteur et la cible afin d'aligner la direction de l'effecteur avec la cible. Cela est répété

pour chaque articulation jusqu'à ce que la position cible soit atteinte. Cet algorithme est simple à implémenter et converge, mais peut parfois donner des postures non réalistes. *FABRIK* est un solveur itératif qui modifie progressivement les positions des articulations de la chaîne pour minimiser la distance entre la cible et l'effecteur final, en opérant en deux étapes : avant et arrière. Dans la phase avant, l'effecteur final est placé sur la cible. Les autres articulations sont mises à jour une par une jusqu'à la racine, en résolvant la distance qui les sépare à l'articulation fille. Le même processus est répété, mais dans l'autre sens : la racine est replacée sur sa position d'origine et les articulations de la chaîne sont tirées en arrière. Les avantages de *FABRIK* sont sa simplicité et sa flexibilité. Il est prouvé que l'algorithme converge toujours lorsque la cible est à atteignable et il peut gérer de nombreux types de contraintes [ACL16].

Métaphores d'édition Dans l'animation traditionnelle par dessin, il est courant que les animateurs commencent à poser en esquissant une version simplifiée du personnage pour se faire une idée rapide de ce à quoi ressemblera la pose. Un ensemble de travaux s'est alors intéressé à la transposition de ce flux de travail à l'animation par ordinateur [Las01] (voir la figure 2.19). Un paradigme intuitif est de demander aux utilisateurs de dessiner en 2D des objets en bâton et de les faire bouger avec des courbes [TBP04 ; Dav+06 ; CON05]. Pour les personnages, une abstraction inspirée par le flux de travail de dessinateurs traditionnels [Bla49 ; LB84] est le concept de ligne d'action (*LdA*). La *LdA* est une ligne simple, unique et lisse du corps du personnage, qui indique la forme esthétique générale de la pose. Il a été démontré que c'est une interface intuitive pour produire des poses expressives, en manipulant le squelette [GCR13 ; Gua+15] ou le maillage du personnage directement [Özt+13]. Ce paradigme est aussi transposé à une interface de réalité virtuelle combinée aux descripteurs de Laban [GRC19]. Le concept est généralisé pour permettre à l'utilisateur de fournir ses propres courbes pour n'importe quelles sous-parties de personnages virtuels [Hah+15].

Figure 2.19 : Différents paradigmes pour l'édition de poses. Gauche : Stick-figures [Dav+06]. Milieu : ligne d'action [GCR13]. Droite : généralisation des lignes d'action [Hah+15].

Animation se basant sur des données

Machine à états et mélange d'animations La machine à états est une technique utilisée pour combiner des animations en fonction de l'état d'un personnage ou d'un objet [LK05 ; KPS13]. Par exemple, un personnage pourrait avoir des animations différentes pour mar-

cher, courir et sauter, et la machine à états permet de passer d'une animation à l'autre en fonction de l'action que le personnage est en train de réaliser. Les transitions se réalisent ici pour un clip complet. Pour rendre l'animation continue et fluide, il faut combiner l'approche à une technique de mélange d'animations (*motion blending*) [PSS02].

Graphe d'animations Plutôt que de raisonner au niveau du clip, les graphes d'animations proposent de raisonner par poses [AF02 ; KGP08] comme illustré sur la figure 2.20. Pour enchaîner des animations, la transition entre deux clips doit se réaliser entre deux poses assez proches. Le graphe d'animation structure une base de données de clips de mouvements en graphe afin de produire des comportements animés sans interruption. Dans ce graphe, chaque nœud représente une pose, chaque arête représente une transition possible entre deux poses du même clip ou de deux clips différents. Parcourir le graphe produit une animation continue unique dans laquelle la notion de clip n'a plus vraiment de sens, seule compte la continuité de l'animation en respectant les contraintes qui sont souvent de suivre une direction ou un chemin. Le coût de recherche dans un graphe de mouvements peut croître exponentiellement avec le nombre d'arêtes. Safonavo *et al.* [SH07] proposent des algorithmes de recherche optimisés en regroupant des poses en méta-nœuds ou alors en pré-calculant des chemins courants.

Les graphes d'animations ne permettent les transitions entre les clips que lorsqu'ils atteignent les nœuds proposant un arc de passage, ce qui peut être contraignant dans les applications demandant de la réactivité, à moins d'avoir une variété de données importantes. Les champs de mouvements (*Motion Field*) [Lee+10] proposent alors de combiner plus de deux poses pour offrir de plus grandes possibilités. L'apprentissage par renforcement est utilisé pour pré-calculer à l'avance quelles combinaisons de poses sont nécessaires pour des scénarii prédéfinis.

Les graphes d'animations sont désormais très répandus dans l'industrie des jeux, mais plutôt comme une machine à états élaborée où le graphe est construit à la main par un animateur. Les approches de construction automatique se sont en effet révélées difficiles à contrôler, ce qui entraîne une latence peu acceptable pour un contrôle interactif. La construction du graphe est alors devenue une tâche fastidieuse pour les animateurs.

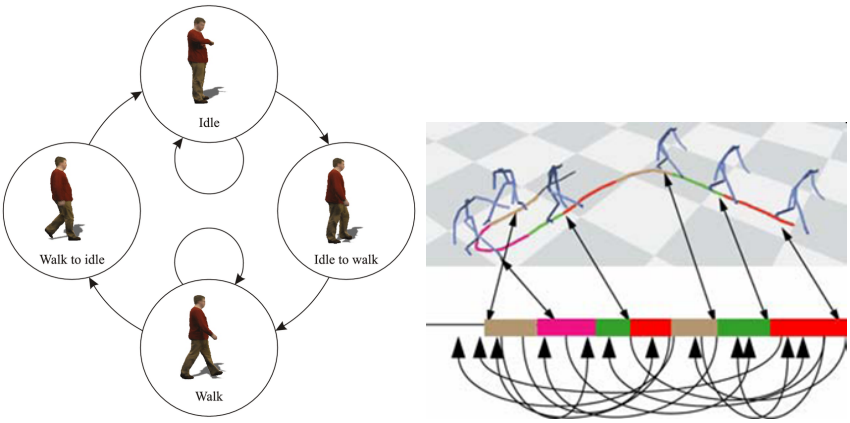
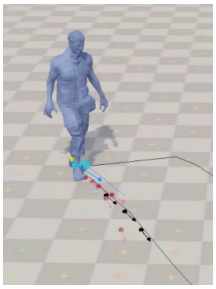


Figure 2.20 : À gauche, les approches à base de machines à états [LK05] organisent une continuité des animations en passant d'un clip à un autre. À droite, le graphe d'animation [AF02] combine une base d'animations en considérant une pose par nœud du graphe.

Apprentissage profond Holden *et al.* [HSK16 ; HKS17] proposent une approche à base de réseaux de neurones capable de fournir une fonction de phase décrivant les pas successifs du personnage. À partir de la pose courante, de la phase et de la direction désirée, le réseau produit la pose suivante. Dans une première version, un auto-encodeur produit la pose du corps suivi d'une optimisation pour garantir les longueurs entre les articulations. Une deuxième version, offrant un contrôleur de locomotion impressionnant, produit la pose avec un réseau entraîné selon les quatre phases de la marche. Cette deuxième approche gère également des sols non plats. Cette idée d'avoir un ensemble de réseaux d'experts se concentrant sur chacun sur les sous-parties, le tout orchestré par un réseau de déclenchement est repris dans d'autres travaux [Zha+18b ; Sta+19 ; Sta+20].

Pour être exhaustif, il est intéressant de citer une catégorie de solutions qui produit les mouvements d'agents articulés plongés dans une simulation physique [Pen+17 ; Pen+18]. Ces méthodes arrivent à simuler des effets liés à la physique comme l'équilibre, car l'optimisation vise le respect de la contrainte de ne pas tomber, en plus de se déplacer dans la direction désirée. Elles se basent souvent sur l'apprentissage par renforcement et demandent de très long temps de calcul. Elles sont très peu contrôlables par un animateur, car chaque nouvelle contrainte demande de modifier la fonction de coût et d'attendre plusieurs heures avant de voir le résultat du contrôleur appris.

Motion matching Plus récemment, une autre méthode appelée *Motion Matching* est apparue dans l'industrie des jeux vidéo [Cla16 ; Zad16] comme une simplification des approches précédentes. À chaque pose, sans pré-calcul, l'algorithme recherche la pose suivante la mieux adaptée dans une base de données de mouvements complète et non structurée. Cette approche supprime la phase pré-calculée. Toute la base de données est conservée en mémoire, et le clip le mieux adapté à chaque étape est sélectionné à travers une recherche de pose la plus proche respectant les contraintes désirées comme la direction de déplacement voulue. La simplicité et la capacité de contrôle ont rapidement attiré l'attention de nombreuses applications de type jeux vidéo [Zin19 ; Abi+22]. L'approche a été combinée à une série de réseaux de neurones, permettant de diviser par dix la taille mémoire occupée par la base de données [Hol+20a] comme le montre la figure 2.21.



	Sélection par recherche de mouvement	Décompresseur	Sélection apprise de mouvement
Animations sources	461.0 Mo	-	-
Caractéristiques de sélection	129.8 Mo	129.8 Mo	-
Caractéristiques complémentaires	-	47.2 Mo	-
Paramètres du Décompresseur	-	2.4 Mo	2.4 Mo
Paramètres du Séquenceur	-	-	1.4 Mo
Paramètres du Projecteur	-	-	13.0 Mo
Total	**590.8 Mo**	**179.4 Mo**	**16.8 Mo**

FIGURE 2.21 : Dans la technique du *motion matching* [Cla16], pour la pose suivante, l'algorithme explore toutes les autres poses possibles dans la base de données et choisit celle respectant les contraintes de direction, de vitesse et de position de pieds. À droite, une version à base de réseaux de neurones [Hol+20a] permet de limiter fortement la taille mémoire occupée.

2.4.2 Édition et synthèse de style

Dans les approches d'édition et de production d'animations décrites jusque-là, la notion d'expression et de style est souvent peu présente directement. En effet, le domaine de l'animation a d'abord cherché à animer les personnages virtuels en laissant le problème de créer des animations expressives à la capture de mouvements et aux animateurs. Dans un cas extrême, on peut demander aux acteurs en studio de capture de réaliser un mouvement pour chaque combinaison de mouvement/style nécessaire à l'application. Cependant, ce processus est très coûteux en temps et en moyen humain. L'autre méthode pour réaliser cette base de données est de demander à différents artistes de créer ou modifier des animations à la main, ce qui est très long également. Cependant, il existe quelques techniques qui assistent l'animateur dans ce processus.

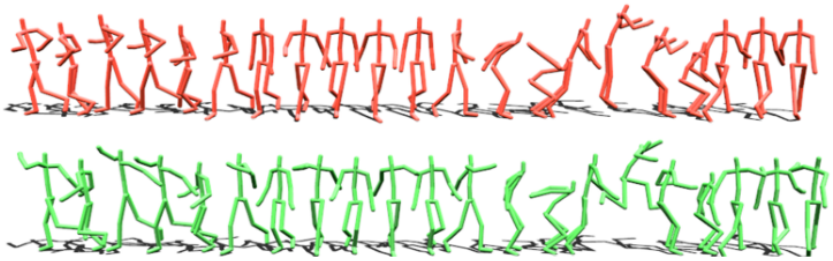


Figure 2.22 : un exemple d'animation dont le style est modifié par l'approche de Xia *et al.* [Xia+15a] qui se base sur des mixtures de modèles régressifs locaux.

Approches statistiques

Pour pouvoir modifier une animation et un style, certaines approches cherchent à construire un modèle statistique de mouvement en faisant ressortir des variations de facteurs [Dia+10]. Un mouvement est alors exprimé comme une combinaison de représentations de base. L'Analyse en Composantes Principales (*ACP*) en est la version la plus représentative. On obtient une représentation plus compacte et générale qui permet d'éditer une pose ou un geste [Gro+04a]. Mais d'autres techniques existent comme celles se basant sur les processus gaussiens pour apprendre une fonction de mapping entre un espace latent à basses dimensions et un espace à hautes dimensions. Certains travaux [HPP05 ; Law04 ; Gro+04b ; WFH07] appliquent ceci sur les poses d'un personnage. En combinant ces modèles avec un système d'*IK*, on peut éditer la pose et le style. Dans le même esprit d'avoir un espace latent représentatif, mais plutôt pour les mouvements complets, d'autres approches se basent sur une régression linéaire multiple [Ma+10 ; Liu+05] ou sur un modèle de Markov caché pour différents individus [BH00 ; WFH07 ; Ma+10]. Ainsi, ces méthodes permettent d'assurer aux animations produites d'avoir une structure de mouvement humain correct, incluant les informations de contact avec l'environnement ce qui permet de réduire les artefacts tels que les glissements des pieds sur le sol. Parmi les méthodes permettant de modifier le style d'une animation de manière robuste sur des données très hétérogènes, c'est-à-dire contenant des actions différentes et éloignées, nous pouvons citer la méthode de Hsu *et al.* [HPP05] ou de Xia *et al.* [Xia+15a]. Ils modélisent les différences de style en

utilisant une série de mixtures de modèles régressifs locaux et peuvent alors gérer des données non labellisées et hétérogènes dont des résultats sont illustrés sur la figure 2.22.

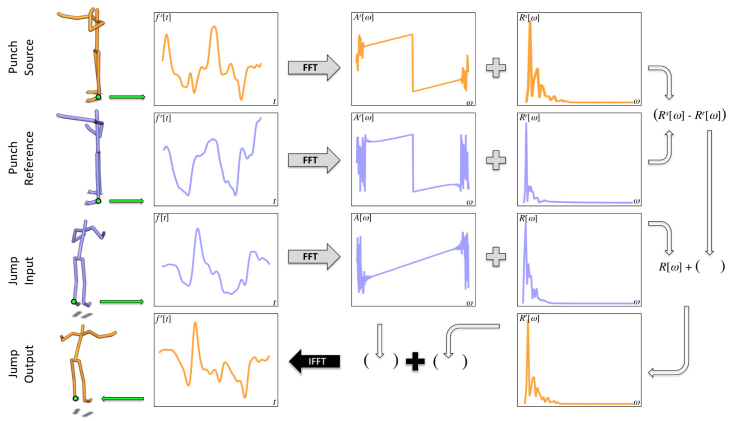


FIGURE 2.23 : L'extraction du résidu est obtenue par soustraction dans le domaine spectral entre le mouvement neutre (Source) et le mouvement contenant l'expression souhaité (Référence). Le résidu est ensuite appliqué sur le mouvement neutre souhaité (Input) et permet d'obtenir le mouvement souhaité avec l'expression choisie (Output). Source : [YM16].

Approches inspirées du traitement du signal

Une autre famille de méthodes permettant de modifier le style est celle traitant les différentes valeurs d'angles pour chaque articulation comme un signal temporel sur lequel différents filtres sont appliqués. Il est possible de décomposer le signal avec un filtrage multi-résolutions ou avec la transformée de Fourier [BW95 ; UAT95], ce qui permet l'édition fréquentielle un peu comme les égaliseurs le permettent avec le son. En mélangeant les coefficients de Fourier, on peut également obtenir des transitions entre des animations de manière fluide et réaliste. En considérant les bonnes fréquences, Breuderlin *et al.* arrivent à extraire "l'humeur" afin de générer des animations riches et variées. De même, Witkin et Popovic [WP95] introduisent un algorithme appelé *motion warping* qui permet de déformer les paramètres des courbes d'animations comme un signal. Amaya *et al.* [ABC96] cherchent à calculer des transformations d'émotions qui vont être ensuite appliquées à des animations existantes afin de modifier l'expression du mouvement. Pour cela, les transformations cherchées capturent la différence entre un mouvement neutre et un mouvement expressif en respectant la vitesse et l'amplitude spatiale du mouvement. Ding *et al.* [Din+14] proposent d'apprendre la relation entre des signaux d'entrée (acoustiques) et les mouvements humains afin de produire automatiquement des animations de rire en temps réel. La méthode proposée par Yumer et Mitra [YM16] permet de transférer le style d'une animation sur différentes actions hétérogènes en utilisant la transformée de Fourier, comme l'illustre la figure 2.23. En calculant la différence dans le domaine spectral entre une animation neutre et une animation stylisée réalisant le même mouvement, ils obtiennent le résidu contenant uniquement la différence entre ces animations, c'est-à-dire le style. En utilisant cette fonction de coût synthétisant le mouvement stylisé, ils peuvent transférer le style sur des animations réalisant des mouvements différents en respectant des contraintes

sur le timing pour pouvoir correctement transférer le résidu sur les articulations effectuant le mouvement.

Approches basées sur l'apprentissage

Plus récemment, les méthodes d'apprentissage profond ont montré une bonne capacité à modifier des animations à partir de grandes bases de données, avec peu de supervision. Par exemple, Holden *et al.* [Hol+15] emploient un auto-encodeur convolutif et montrent que l'interpolation dans l'espace latent permet, contrairement à l'interpolation des coordonnées ou des quaternions, de produire des mouvements réalistes. Le potentiel de cette architecture est amélioré dans [HSK16 ; Hol+17], en ajoutant des couches afin de prendre en compte le contact des pieds avec le sol. En s'inspirant du transfert de style pour les images [GEB16 ; Jin+19], Holden *et al.* couplent cet auto-encodeur à la matrice de Gram pour transférer le style d'une animation à une autre [Hol+17]. Manocha *et al.* travaillent spécifiquement sur la démarche. Ils dérivent leurs travaux en reconnaissance de démarches [Ran+22 ; Bha+20] utilisant les réseaux convolutionnels pour générer des comportements émotionnels d'agents virtuels en utilisant des caractéristiques expressives de la démarche et du regard [Ran+19 ; Ran+22]. Leurs approches vision/synthèse trouvent des échos particuliers avec nos travaux. De manière plus générale, d'autres types de réseaux sont appliqués pour transférer un style entre deux animations avec un exemple sur la figure 2.24 : les *GAN*[Wan+18b ; Wan+20a], des ensembles de réseaux experts travaillant sur des sous-parties du corps [Smi+19 ; Abe+20c], des réseaux neuronaux probabilistes génératifs [Wen+21], les réseaux d'attentions [Bha+20], des réseaux combinant texte et animation [Bha+21], etc. Ces techniques obtiennent des résultats intéressants, mais leur défaut majeur est de ne produire en sortie que des variations de gestes issues des bases peu fournies et surtout de n'autoriser aucun contrôle par un animateur. Le réseau apprend et applique, sans fournir de paramètres compréhensibles par un humain. Les approches de génération de visages [Tov+21 ; Xia+22] ont de l'avance sur ce point, car elles ont déjà franchi le pas en proposant des réseaux capables de se déplacer dans un espace latent en offrant des paramètres sur les caractéristiques du visage.

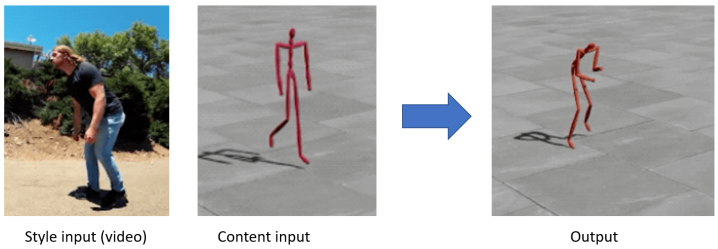


Figure 2.24 : Aberman *et al.* [Abe+20c] arrivent à transférer un style d'une vidéo de personne vers un personnage.

2.4.3 Bilan de l'existant en synthèse d'animations

En analysant les différentes méthodes en synthèse d'animations de personnages virtuels, on peut constater qu'une grande majorité des techniques se basent sur des données grâce

à la capture de mouvements. La recherche a d'abord tenté de combiner les animations afin de répondre aux applications qui veulent au minimum diriger un personnage dans une direction donnée sur des terrains variés. Il y a récemment une évolution forte vers l'utilisation de plus en plus d'apprentissages automatiques et plus spécifiquement d'apprentissage profond. Cependant, représenter l'espace des animations dans un espace latent produit par des réseaux de neurones ne garantit pas facilement des paramètres compréhensibles et éditables par un humain. Le domaine de l'apprentissage ou de la vision cherchent souvent des techniques entièrement automatiques. Alors qu'en animation, les applications demandent souvent qu'un artiste puisse s'exprimer avec des outils lui laissant une place. Il nous semble alors que les approches pertinentes essaient de mixer des solutions expertes, basées sur des paramètres définis par des spécialistes du mouvement, avec des approches à base d'apprentissages automatiques. Les réseaux sont souvent de petites tailles et se concentrent sur une petite partie du problème. Cette philosophie de diviser pour régner permet aussi d'inclure l'animateur dans les résultats intermédiaires.

Concernant plus spécifiquement la synthèse ou l'édition de styles, le domaine devient encore plus restreint. Il existe des approches capables de transférer automatiquement un style d'une animation vers une autre, même en temps réel avec comme source une capture de mouvements réalisée par une simple webcam. Cependant, les espaces de représentation restent très peu adaptés à fournir des outils à un animateur qui dispose d'un savoir-faire pointu. À notre connaissance, il n'existe aucune approche récente capable d'offrir des paramètres experts comme ceux issus de la reconnaissance automatique pour éditer une animation. Il reste donc de la place pour explorer des méthodes s'inspirant de l'apprentissage automatique et de la vision, tout en incluant le spécialiste et en restant raisonnable sur les quantités de données nécessaires.

3 Analyse et reconnaissance des expressions du visage

La communication sous toutes ses formes, verbale ou non verbale, est importante pour accomplir diverses tâches entre les humains et joue un rôle important dans la vie quotidienne. L'expression faciale est la forme la plus efficace de communication non verbale et elle fournit un indice fort sur l'état émotionnel, l'état d'esprit et l'intention. Pour la communauté de la vision par ordinateur, la reconnaissance automatique des expressions faciales en temps réel avec une grande fiabilité est une tâche difficile qui anime les chercheurs depuis longtemps [SI92]. La variabilité de la pose, de l'éclairage, les occultations, les spécificités de chacun et de chaque culture dans la manière d'exprimer ses émotions sont quelques-uns des paramètres qui rendent cette tâche difficile. Après plusieurs décennies de travaux, l'état de l'art montre que de nombreuses méthodes atteignent des résultats très bons lorsque les conditions d'acquisitions sont maîtrisées. Le domaine s'attaque alors à des caractéristiques particulières, à des spécificités afin de se spécialiser et de rendre les approches plus robustes à tous les types de cas. En effet, les conditions peuvent varier en fonction de l'environnement (*i.e.* éclairage direct et diffus), la personne (*i.e.* âge, genre, type, etc.), le point de vue (*i.e.* face à la caméra, occultation, lunettes, etc.), et les outils de captures (*i.e.* qualité de l'image). Ce domaine de recherche avance constamment. Les réseaux de neurones ont beaucoup apporté pour augmenter la capacité des approches à généraliser la reconnaissance à des situations comportant moins de contrôle. Le schéma classique de reconnaissance consiste à extraire des caractéristiques pertinentes des images et à les fournir à un classifieur qui produit le label après un entraînement (voir la figure 3.1). Nous abordons deux problématiques particulières. La première est de proposer une approche temps réel, robuste aux faibles résolutions des images d'entrée en s'inspirant du système visuel humain. La deuxième concerne les visages d'enfants et les expressions spontanées, deux caractéristiques très peu abordées dans l'état de l'art.

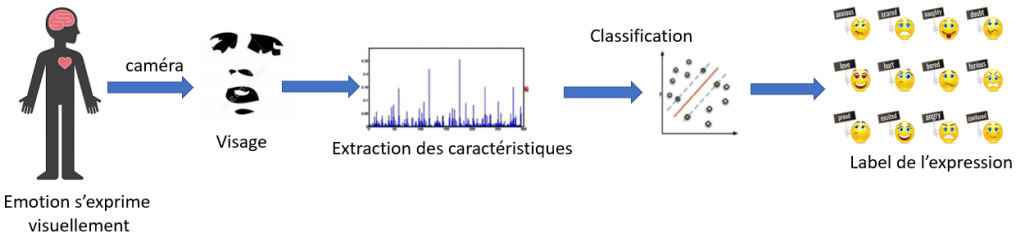


Figure 3.1 : Dans ce chapitre, des contributions sur la reconnaissance d'expressions faciales sont proposées en s'intéressant aux spécificités des images à faible résolution et des expressions d'enfants.

3.1 Reconnaissance d'expressions faciales inspirée par le système visuel humain

Le Système Visuel Humain (SVH) décode et analyse les expressions faciales en temps réel malgré des ressources neuronales relativement limitées. Pour expliquer cette performance, il est proposé que seules certaines entrées visuelles soient sélectionnées en considérant les "régions saillantes" [Zha06], où "saillant" signifie les plus visibles ou les plus importantes. Les images d'entrée à faible résolution rendent cette tâche d'identification encore difficile. Les vidéoconférences dans de grandes salles, les jeux en ligne, l'analyse de films sont quelques-unes des applications qui gagneraient en fonctionnalités avec des systèmes de reconnaissance des expressions faciales fonctionnant correctement sur des images de visages à faible résolution.

À partir de l'étude de l'état de l'art des descripteurs de texture et en s'inspirant du système visuel humain, il est possible de cibler avec succès une spécificité particulière dans le domaine de la reconnaissance des expressions du visage : les images de faibles résolutions. Les travaux présentés ici sont détaillés dans la thèse de Rizwan A. Khan [Kha13]. La première piste proposée est un nouveau type de descripteur multi-résolutions appelé pyramide de motifs binaires locaux (*PLBP*). L'objectif est de conserver la bonne propriété des *LBP* (*Local Binary Pattern*) [ZP09] de s'adapter aux changements d'illumination, au traitement des images de faible résolution. La deuxième piste étudiée est de ne traiter algorithmiquement que les régions faciales saillantes en se basant sur une étude du système visuel humain. Traiter seulement certaines zones du visage permet de réduire la longueur du vecteur de caractéristiques et donc de diminuer les temps de calcul.

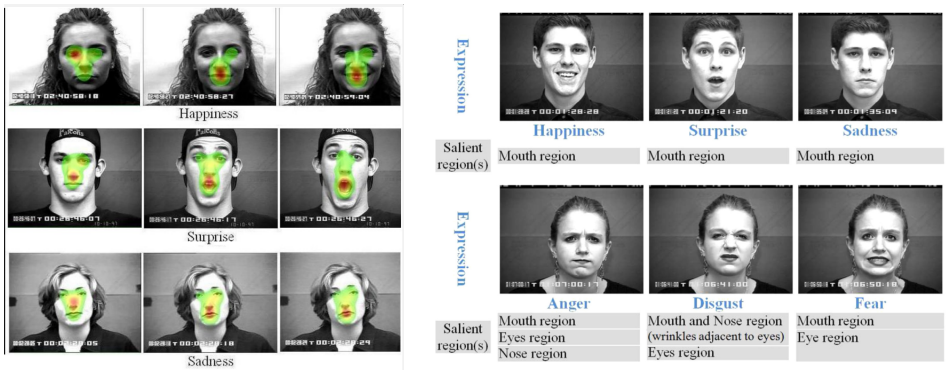


Figure 3.2 : Nous avons conduit une étude expérimentale du système visuel humain afin de déterminer si des régions d'un visage exprimant une émotion étaient observées en priorité, ceci pour chacune des six émotions de base. Ces régions saillantes sont utilisées pour optimiser la reconnaissance d'expressions. À gauche, des exemples de zones observées. À droite, la liste des régions observées prioritairement pour chaque expression. Source : [Kha+13b]

3.1.1 Régions saillantes

Nous avons réalisé une étude expérimentale psychovisuelle qui a enregistré les mouvements oculaires d'observateurs humains dans des conditions d'observation de visages

exprimant une émotion. Une analyse de ces données indique quelles parties du visage sont saillantes pour chaque expression spécifique, comme le montre la figure 3.2. Les mouvements oculaires de quinze observateurs humains ont été enregistrés à l'aide d'un traqueur de regard (système Eyelink II de SR Research [Eye]). Les sujets ont regardé une sélection de 54 vidéos choisies dans la base de données étendue Cohn-Kanade (CK+) [Luc+10], montrant l'une des six expressions faciales universelles. Les observateurs sont des hommes et des femmes âgés de 20 à 45 ans, ayant une vision normale ou corrigée.

Les conclusions tirées de cette étude suggèrent que pour certaines expressions (la joie, la tristesse et la surprise), une seule région du visage est saillante, tandis que pour les autres expressions, deux régions du visage sont saillantes ou contiennent la plupart des informations discriminantes. La tâche d'analyse et de reconnaissance des expressions peut alors être effectuée de manière plus efficace (voir la figure 3.3), si seules les régions perceptives saillantes sont sélectionnées pour le traitement, comme cela se produit dans le système visuel humain. En ne traitant que les régions perceptives saillantes, il est possible de réduire la dimension du vecteur de caractéristiques. Cette réduction de la longueur du vecteur de caractéristiques rend l'approche appropriée pour les applications en temps réel en raison de la complexité de calcul minimisée et ceci reste vrai même avec des approches récentes à base de réseaux de neurones.

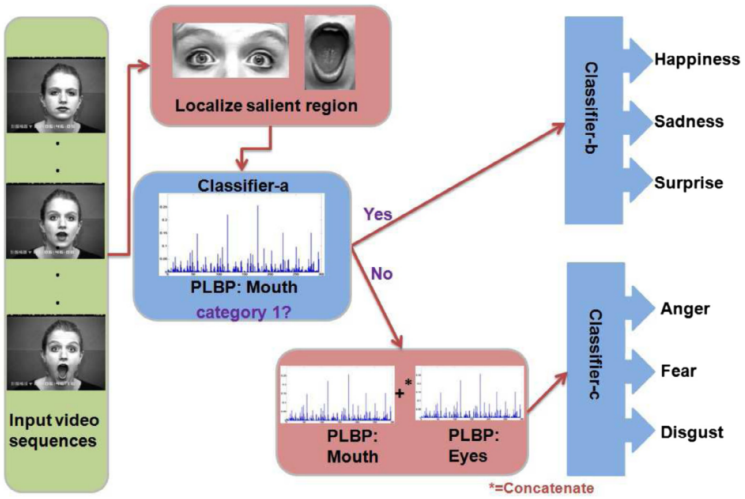


Figure 3.3 : Une analyse de la bouche permet de reconnaître la joie, la tristesse, la surprise ou de décider s'il faut ajouter les yeux au processus pour reconnaître la peur, le dégoût ou la colère. Source : [Kha+13b]

3.1.2 Motifs binaires locaux multi-résolutions

L'analyse de l'état de l'art (section 2.2.2) montre que la reconnaissance des expressions du visage se base souvent sur la texture du visage extraite de la vidéo. Cette texture comporte des motifs dont certains ont une forme bien précise comme les pourtours des yeux, de la bouche et du nez, complété par toutes les rides de peau. Il a été montré que les rides aident la perception des expressions [CBM09]. Dans nos travaux présentés en 5.1, une carte de normales est capturée sur des images de visages et transférée pour une animation 3D. Pour

la reconnaissance d'expressions faciales, l'idée d'explorer ces textures de visage semble donc une bonne piste. Il existe une approche assez courante en analyse de texture : les caractéristiques *LBP* (*Local Binary Pattern*) [OPM02]. La propriété la plus importante des caractéristiques *LBP* est leur tolérance aux changements d'illumination et leur simplicité de calcul. L'algorithme étiquette les pixels d'une image avec un seuil calculé sur le voisinage 3×3 de chaque pixel avec la valeur centrale et en considérant le résultat comme un nombre binaire. Ensuite, l'histogramme des étiquettes peut être utilisé comme descripteur de texture.

Une approche multi-résolutions est classique pour caractériser des informations de différentes tailles. Nous introduisons un traitement multi-résolutions appelé pyramide de motifs binaires locaux multi-résolutions (*Pyramid Local Binary Pattern - PLBP*) pour l'analyse des caractéristiques faciales (voir la figure 3.4). Les *LBP* représentent les dispositions spatiales de la texture. La disposition spatiale est acquise en divisant l'image en régions à des résolutions multiples. Si le niveau le plus grossier est utilisé uniquement, le descripteur revient à un histogramme *LBP* classique. En comparaison avec l'approche *LBP* originale travaillant sur les textures [OPM02], le descripteur proposé ici sélectionne les échantillons de manière plus uniforme, alors que le *LBP* d'Ojala *et al.* prend des échantillons centrés autour d'un point, ce qui conduit à manquer certaines informations dans le cas d'un visage (qui est différent d'une texture répétitive). Ce descripteur multi-résolutions est une extension simple et efficace en termes de calcul du descripteur de texture *LBP*.

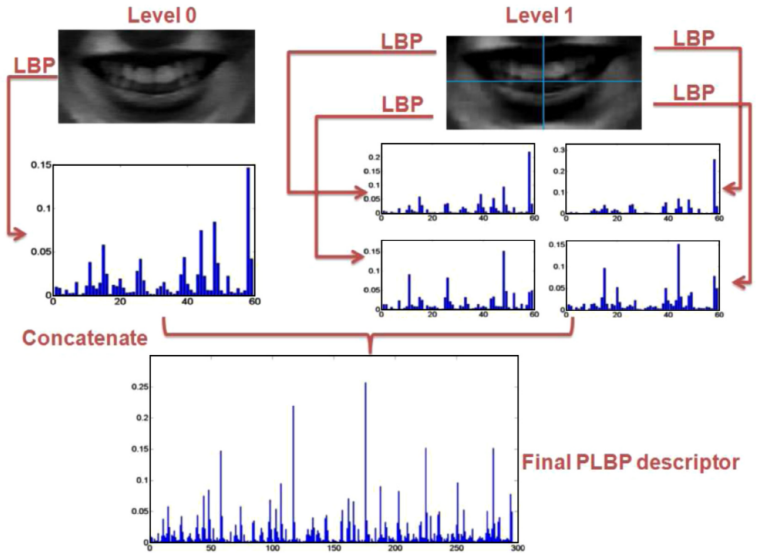


Figure 3.4 : Nous introduisons le descripteur *PLBP* (pyramide de motifs binaires locaux multi-résolutions) pour la reconnaissance d'expressions faciales. La zone du visage traitée est découpée en sous-régions pour former un descripteur multi-résolutions. Source : [Kha+13b]

3.1.3 Reconnaissances des six expressions universelles et de la douleur

Les deux contributions proposées, à savoir les nouveaux descripteurs *PLBP* et le fait de se concentrer uniquement sur la ou les régions faciales visuellement saillantes, sont testées sur les bases de données classiques du domaine : Cohn–Kanade, Cohn–Kanade Plus [Coh+03], et *FG-NET FEED* [Wal+06]. Ces bases sont décrites en détail dans la section 2.2.3 et comportent les six expressions basiques : la colère, le dégoût, la peur, la joie, la tristesse et la surprise. Plusieurs méthodes courantes de classification sont testées (SVM, Random Forest, arbres de décision) avec des taux de bonne reconnaissance supérieurs à 90% pour des images de résolutions natives. Mais l'avantage de cette approche est qu'elle est invariante par rapport à l'illumination et surtout extrêmement fiable sur des images de faible résolution comme illustré par la figure 3.5.

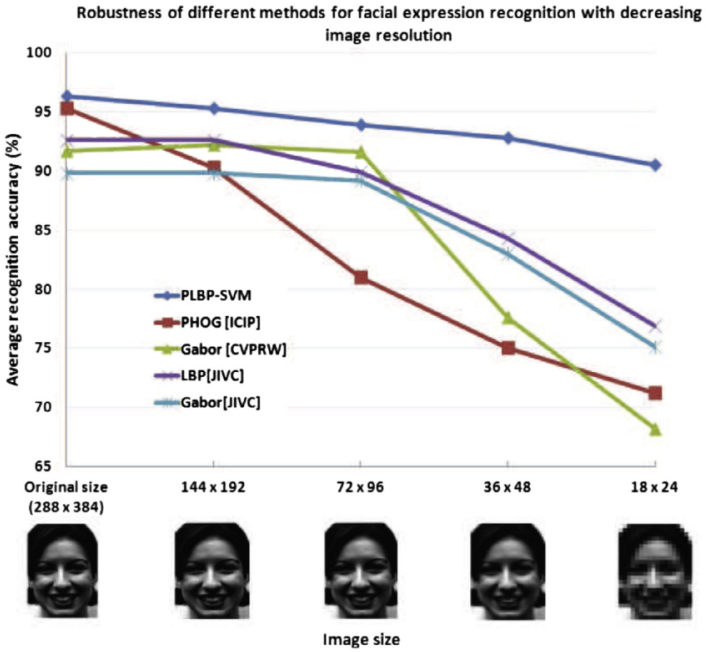


Figure 3.5 : Comparaison de robustesse aux faibles résolutions des images en entrées pour différentes méthodes. Notre approche *PLBP-SVM* garde une bonne performance même sur des images de faible résolution. Source : [Kha+13b]

Les descripteurs *PLBP* sont également testés sur une base de données d'expressions faciales de douleur [Luc+11] avec des taux de bonne reconnaissance de 96%. De nombreuses applications médicales peuvent profiter de cette technique pour aider aux soins. L'optimisation consistant à ne travailler que sur certaines régions saillantes n'est pas utilisée ici, car elle repose sur l'étude du système visuel humain et les six expressions de base, mais la douleur n'en fait pas partie. De plus amples détails sont disponibles dans la publication [Kha+13a].

3.1.4 Mise en relation avec les approches actuelles

Avec le recul, il est raisonnable d'affirmer que l'idée de découper le visage en trois zones (yeux, nez, bouche) et d'analyser uniquement certaines zones saillantes comme semble le faire le système visuel humain est une voie importante à explorer. Il est possible que les réseaux de neurones profonds trouvent des raccourcis plus ou moins similaires, mais ceci serait au prix d'une quantité de données nettement supérieure. Les bases de données actuelles comportent environ un million d'images. Bien sûr, de l'information est présente dans ces données, mais rien ne garantit qu'il soit facile de mettre au point un réseau capable de la trouver sans faire du sur-apprentissage. D'un autre côté, si un expert peut proposer des pistes de transformations des données afin d'extraire une information plus concentrée, il est toujours intéressant de se baser dessus, même si ensuite ces données sont traitées par un réseau. L'étude du système visuel humain a permis de fournir ce concentré d'informations. Il serait sûrement intéressant d'injecter ces points dans les réseaux de neurones actuels pour en diminuer la taille, les temps d'apprentissage et peut-être pour voir les réseaux faire émerger d'autres éléments en parallèle. D'autant plus que les approches de *"few shot learning"* semblent avoir un avenir intéressant [Wan+20b] dans la communauté de l'apprentissage machine et de la vision par ordinateur.

3.2 Expressions faciales de visages d'enfants

L'étude de l'état de l'art sur la reconnaissance d'expressions de visages présentée dans la section 2.2 et dans l'article de Li *et al.* [LD22] montre que le domaine passe progressivement des conditions contrôlées en laboratoire aux conditions dites *"in the wild"* où tous types de situations peuvent se produire : éclairage, occultations, qualités des images, orientation de la tête, identité et ethnie de la personne, etc. Dans la section précédente, une approche experte a été proposée afin de pouvoir traiter une de ces spécificités : les images de faibles résolutions. Dans la stratégie d'étudier une seule spécificité à la fois, les expressions d'enfants sont abordées dans cette section, en offrant une nouvelle approche de reconnaissance couplée à une nouvelle base de données d'expressions spontanées de visages d'enfants. Cette nouvelle base de données de vidéo d'expressions d'enfants est nommée *LIRIS-CSE* pour *LIRIS Children Spontaneous facial Expression* [Kha+20].

3.2.1 Pourquoi une nouvelle base de données ?

Le succès récent des techniques d'apprentissage profond dans de nombreux domaines repose en partie sur la disponibilité de grandes quantités de données. De plus, il est important que les bases de données représentent toutes les spécificités de la vie courante. L'analyse des bases de données existantes présentées dans la section 2.2.3 montre différents biais. En effet, il existe un certain nombre de bases de données de référence accessibles au public, comportant des visages de personne adultes jouant les six émotions de base. Malheureusement, une grande majorité de ces bases de données propose des expressions actées. Le faible nombre de bases de données avec des expressions spontanées [VP10] est un frein au développement de techniques de reconnaissance d'expressions faciales

Figure 3.6 : Exemples d'apparition d'une expression. La première ligne montre l'apparition du "Dégoût", la deuxième ligne de la "tristesse", et la troisième de la "Joie".

applicables à une tâche réelle. En effet, il est prouvé que les expressions faciales spontanées diffèrent des expressions posées [Bar+06].

Toutes ces bases de données ne contiennent souvent que des images statiques, tellement contraintes qu'elles sont assez proches de photos d'identité officielles avec l'expression poussée à son intensité maximale. Selon une étude menée par le psychologue Bassili [Bas79], il a été prouvé que le mouvement des muscles faciaux est fondamental pour la reconnaissance des expressions faciales. Il a également conclu qu'une personne peut reconnaître des expressions de manière plus robuste à partir d'un clip vidéo qu'à partir d'une photo, surtout quand celle-ci ressemble à une photo d'identité officielle.

De plus, Charlesworth *et al.* montrent que les expressions d'enfants [CK73] diffèrent des expressions d'adultes en étant plus accentuées, mais parfois plus ambiguës. Par exemple, il n'est pas rare qu'un enfant éclate de rire quand l'intensité de la surprise est trop forte. Une nouvelle base de données nommée *LIRIS-CSE* est proposée à la communauté tentant une avancée sur chacun de ces points : des données d'expressions spontanées ciblant la spécificité des visages d'enfants dans des vidéos. Cette catégorie de personnes est très peu traitée par les algorithmes et le fait d'obtenir des expressions spontanées évite les biais des expressions actées afin de faire un pas dans le traitement *"in the wild"* du problème.

3.2.2 Constitution de la base de données

La première étape de la création de la base de données sur les expressions spontanées d'enfants a été la sélection de stimuli visuels susceptibles de leur provoquer des émotions. Pour des raisons éthiques et compte tenu du jeune âge des enfants, les stimuli ont été soigneusement sélectionnés et nombreux ont été supprimés pour ne pas avoir un impact négatif à long terme sur les enfants. Par exemple, aucun clip inducteur d'émotions pour l'expression négative de la "colère" n'ont été inclus et très peu de clips pour induire l'émotion de la "peur" et "tristesse" sont présents. Valstar *et al.* [VP10] ont procédé de la même manière auparavant. Pour ces raisons, la base de données proposée contient plus de clips d'expressions positives que d'expressions négatives. Bien qu'il n'y ait pas de clips

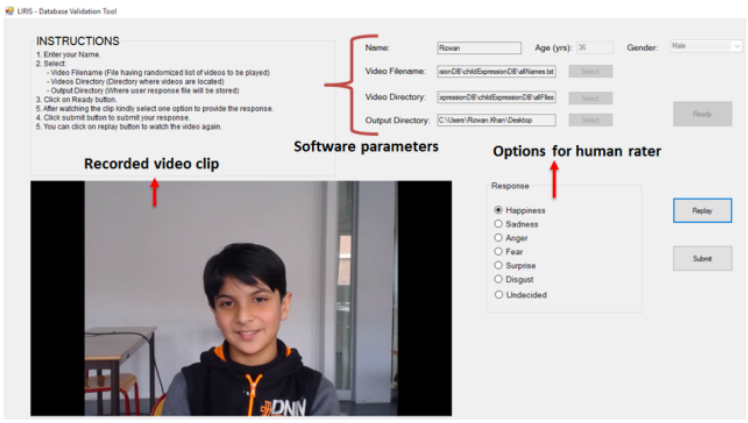


FIGURE 3.7 : Pour la validation de l'expression de chaque séquence, un logiciel permet d'accélérer les traitements pour les évaluateurs humains.

inducteurs d'émotion pour l'expression "colère" Widen *et al.* ont montré que les jeunes enfants utilisent indifféremment les expressions de "dégoût" et de "colère" [WR13].

La sélection comporte uniquement des dessins animés, des extraits de films ou des petits clips vidéo d'enfants faisant des actions amusantes. Les stimuli vidéos peuvent évoquer des émotions pendant une plus longue durée, cela a permis d'enregistrer et repérer les expressions faciales des enfants. Ces stimuli comportent du son pour un meilleur sentiment d'immersion afin d'aider à induire des émotions de manière robuste. Le détail des stimuli est disponible dans la publication [Kha+19a]. Inspiré par Li *et al.* [Li+13], l'enregistrement des vidéos des visages d'enfants a été effectué avec une webcam à haute fréquence, montée sur le haut de l'ordinateur portable avec un haut-parleur et le visage à une distance de 50 cm.

La base de données contient cinq expressions spontanées universelles avec 12 enfants d'origines ethniques diverses (cinq filles, sept garçons), âgés de 6 à 12 ans, avec un âge moyen de 7,3 ans. 60% des enregistrements ont été effectués en classe ou en laboratoire et 40% des clips de la base de données ont été enregistrés à la maison. La figure 3.6 montre trois exemples de séquence temporelle.

Après avoir enregistré la vidéo pour chaque enfant, un post-traitement a été réalisé afin de supprimer toutes les parties inutiles, généralement au début et à la fin de la séquence. Comme les stimuli vidéo regardés par les enfants comportent une combinaison de différentes vidéos émotionnelles, les vidéos enregistrées contiennent également tout un spectre d'expressions en une seule vidéo. Une segmentation manuelle de chaque clip en petits morceaux est réalisée de telle sorte que chaque clip vidéo montre une unique expression prononcée. La base de données a été validée par 22 évaluateurs humains âgés de 18 à 40 ans. Pour cette validation, un logiciel spécialement développé montre aléatoirement une des vidéos segmentées, puis enregistre le choix de l'étiquette d'expression par l'évaluateur humain (voir figure 3.7). Le nombre total de petits clips vidéo, chacun contenant une expression spécifique, présents dans la base de données finale est de 208.

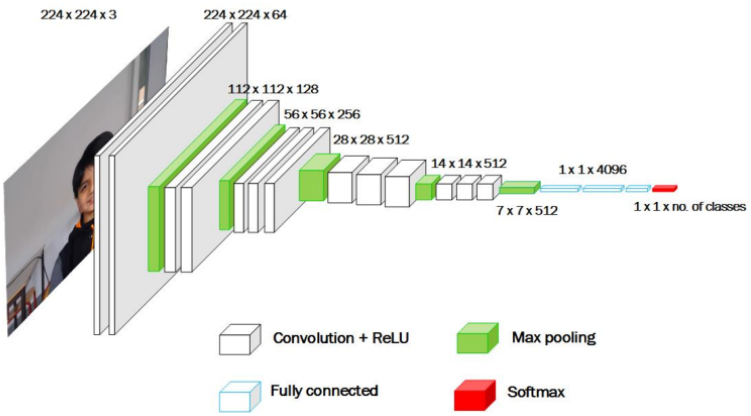


FIGURE 3.8 : Une illustration de l'architecture *VGG16* utilisée pour le transfert d'apprentissage. Source : [Kha+19a]

3.2.3 Transfert d'apprentissage

En général, les réseaux neuronaux convolutifs (*CNN*) nécessitent une quantité d'images importante pour apprendre un concept [LKF10 ; HBW15 ; Alo+18], ce qui peut limiter leur usage. Une piste pour limiter les effets de ce verrou est d'utiliser la technique d'apprentissage par transfert [PY09 ; Ng+15]. L'apprentissage par transfert est une approche automatique qui se concentre sur la capacité à appliquer des connaissances pertinentes issues d'expériences d'apprentissage antérieures à un problème différent, mais connexe. Ce paradigme est appliqué ici pour entraîner un réseau de reconnaissance des expressions sur la base de données proposée (*LIRIS-CSE*), car la taille de cette base de données n'est pas suffisamment importante pour entraîner de manière robuste toutes les couches du réseau depuis zéro. Le réseau *VGG16* à 16 couches [SZ14] est un réseau convolutif profond pré-entraîné par le *Visual Geometry Group (VGG)* de l'Université d'Oxford (voir la figure 3.8). Il est un bon choix pour être spécialisé, car il donne des résultats performants et robustes pour la reconnaissance d'objets issus de la base *ImageNet* [Den+09], spécifiquement conçue pour la reconnaissance d'objets.

La dernière couche entièrement connectée du modèle pré-entraîné *VGG16* a été remplacée par une couche dense ayant cinq sorties. Cela donne 5005 paramètres entraînables. Le nombre de sorties de la dernière couche dense correspond au nombre de classes à reconnaître. Dans ce réseau modifié, il y a cinq sorties pour les cinq classes des expressions à reconnaître. En effet, sur les six expressions universelles, l'expression "colère" n'a pas été incluse, car il n'y a pas de clip pour la "colère" comme expliqué dans la section 3.2.2. La base de données proposée est constituée de vidéos, mais pour cette expérience, des images sont extraites des vidéos. Nous avons utilisé 80% des images pour l'apprentissage et le reste pour la validation. Le réseau *CNN* atteint une précision moyenne de 75% pour la base de données proposée sur les cinq expressions. La figure 3.9 donne la matrice de confusion.

Figure 3.9 : Matrice de confusion. Les lignes correspondent aux labels induits automatiquement. Les colonnes correspondent aux labels donnés par les évaluateurs humains. La diagonale représente l'accord entre les labels induits et les labels des évaluateurs humains. Source : [Kha+19a]

3.3 Conclusion

Ce chapitre décrit deux contributions à la problématique de la reconnaissance d'expressions faciales. La première approche propose de traiter la reconnaissance automatique des expressions faciales en se basant sur une étude initiale de la vision humaine. Cette étude a permis d'extraire les régions faciales saillantes, ce qui permet de proposer un algorithme où les descripteurs sont calculés uniquement sur un minimum de régions. Nous introduisons également des descripteurs de texture multi-résolutions (*PLBP*) qui améliorent la capacité de discrimination par rapport à l'état de l'art. L'ensemble de l'approche est rapide et robuste. En effet, elle fonctionne sur des expressions spontanées, reste fiable sur des images de faibles résolutions et est très rapide à calculer même sur une machine à la puissance limitée. La deuxième contribution de ce chapitre est l'apport d'une nouvelle base de données d'expressions spontanées d'enfants. Elle est disponible à la communauté scientifique depuis janvier 2021. Début 2023, elle totalise un peu plus de trois cents scientifiques enregistrés pour le téléchargement. Proposer des expressions d'enfants spontanées est un aspect innovant comparé aux bases de données existantes. Nous proposons également une approche à base de transfert capable de spécialiser un réseau pour reconnaître automatiquement l'expression avec un taux de bonnes reconnaissances de 75%.

Publications liées

- Alexandre Meyer, Hector M Briceno et Saida Bouakaz. "User-guided shape from shading to reconstruct fine details from a single photograph". In : *Asian Conference on Computer Vision*. Springer. 2007, p. 738-747
- Rizwan Ahmed Khan, Alexandre Meyer, Hubert Konik et Saida Bouakaz. "Exploring human visual system : study to aid the development of automatic facial expression recognition framework". In : *2012 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*. IEEE. 2012, p. 49-54

- Rizwan Ahmed Khan, Alexandre Meyer, Hubert Konik et Saida Bouakaz. "Human Vision Inspired Framework for Facial Expressions Recognition". In : *2012 19th IEEE International Conference on Image Processing*. Sept. 2012, p. 2593-2596
- Rizwan Ahmed Khan, Alexandre Meyer, Hubert Konik et Saida Bouakaz. "Framework for reliable, real-time facial expression recognition for low resolution images". In : *Pattern Recognition Letters* 34.10 (2013), p. 1159-1168
- Rizwan Ahmed Khan, Alexandre Meyer, Hubert Konik et Saïda Bouakaz. "Pain detection through shape and appearance features". In : *Proceedings of the 2013 IEEE International Conference on Multimedia and Expo, ICME 2013, San Jose, CA, USA, July 15-19, 2013*. IEEE Computer Society, 2013, p. 1-6
- Rizwan Ahmed Khan, Alexandre Meyer et Saida Bouakaz. "Automatic affect analysis : from children to adults". In : *International Symposium on Visual Computing*. Springer. 2015, p. 304-313
- Rizwan Ahmed Khan, Alexandre Meyer, Hubert Konik et Saida Bouakaz. "Saliency-based framework for facial expression recognition". In : *Frontiers of Computer Science* 13.1 (2019), p. 183-198
- Rizwan Ahmed Khan, Arthur Crenn, Alexandre Meyer et Saida Bouakaz. "A novel database of children's spontaneous facial expressions (LIRIS-CSE)". In : *Image and Vision Computing* 83-84 (2019), p. 61-69

Thèses liées

- Rizwan A. Khan. "Expression recognition from videos in uncontrolled environment". Theses. Université Claude Bernard Lyon 1, LIRIS, 2013
- Arthur Crenn. "Capture et transfert d'expression de visages d'enfants pour l'interaction avec des mondes virtuels". Theses. Université Claude Bernard Lyon 1, LIRIS, 2019

La base de données *LIRIS-CSE*

- R. A. Khan, A. Crenn, A. Meyer et S. Bouakaz. *LIRIS-CSE Children Spontaneous Facial Expression Video Database*. 2020. URL : `https://childrenfacialexpression.projet.liris.cnrs.fr`

4 Analyse et reconnaissance d'expressions corporelles

Table des matières du chapitre

L'analyse et la reconnaissance des expressions corporelles sont un axe important de nos travaux (voir la figure 4.1). L'objectif est double : comprendre les phénomènes et éléments qui sont les sources de la perception des expressions afin de proposer des approches automatiques de reconnaissance ; dans l'idéal, formaliser ces phénomènes en descripteurs compréhensibles et intuitifs afin de pouvoir les utiliser en synthèse d'animations. Ce thème est abordé au chapitre 5. Un mouvement est couramment représenté en informatique par une succession de poses. Chaque pose est représentée par la position 3D de chaque articulation ou l'angle entre deux articulations successives du corps. Notre objectif principal est de comprendre comment l'émotion ou le style sont exprimés dans ces données en étant capables d'en calculer des caractéristiques. Ces caractéristiques sont validées en les utilisant pour faire de la reconnaissance automatique. Elles peuvent également servir pour l'édition d'animations.

Dans ce contexte, des descripteurs qualifiés d'experts sont introduits dans la section 4.1. Le terme expert provient du constat que ces descripteurs découlent des intuitions induites par la lecture des études réalisées par les experts souvent issues du domaine de la psychologie, des arts ou de l'informatique. Dans la section 4.2, une approche séparant l'action réalisée de l'expression en générant automatiquement un geste neutre est présentée. En 4.3, une approche purement "apprentissage profond" est introduite. Cependant, il est intéressant de voir si des descripteurs experts poussés à leur limite peuvent rivaliser avec les approches moins explicables que sont les réseaux de neurones. Dans la section 4.4, l'analyse des études psychologiques est reprise en associant des valeurs quantitatives aux descripteurs experts. Enfin, dans la section 4.5, une comparaison de ces approches est réalisée en confrontant les taux de bonnes reconnaissances associés à des critères d'explicabilités et de complexités de mise en œuvre.

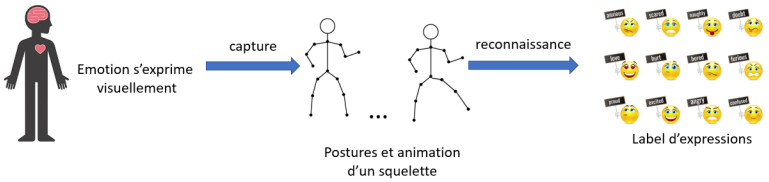


Figure 4.1 : Dans ce chapitre, seules les expressions corporelles et leur reconnaissance sont abordées.

4.1 Descripteurs experts

Une approche basée sur un ensemble de descripteurs qualifiés d'experts est un bon moyen de comprendre les rouages de la problématique de la reconnaissance d'expressions. Des experts (psychologues, chorégraphes, chercheurs en vision, etc.) de différents domaines ont proposé des descripteurs dans leurs publications. Cette section propose une première série. Dans la section 4.4, une analyse plus approfondie est réalisée, en les combinant à des valeurs numériques. Dans cette première version, les descripteurs sont classifiés en trois catégories : géométriques, physiques et fréquentiels. La figure 4.2 décrit le schéma de classification avec ces trois catégories et leur agrégation en méta-descripteurs. Les descripteurs sont calculés puis appris par un classifieur sur des données supervisées. Une fois entraîné, le

classifieur peut alors produire le label de l'expression à partir de caractéristiques jamais observées issues d'une nouvelle animation. Les descripteurs géométriques travaillent sur la posture uniquement et donnent des indications sur la géométrie du corps en considérant certaines distances ou certaines aires de triangles formés par des ensembles d'articulations. Les descripteurs physiques calculent des données issues de la mécanique du mouvement comme la vitesse ou l'accélération relatives de certaines articulations par rapport à d'autres. Les descripteurs fréquentiels se basent sur l'analyse de Fourier appliquée à certaines articulations donnant l'amplitude et la phase. Les caractéristiques de bas niveau sont calculées par fenêtre de temps glissante. Puis, des méta-caractéristiques sont calculées en utilisant la moyenne et l'écart type pour chaque caractéristique de bas niveau sur l'animation complète. Comme les méta-caractéristiques sont indépendantes du temps, l'étape complexe de synchronisation de deux animations est évitée. Les valeurs des méta-caractéristiques sont transmises au classificateur qui fournit le label de l'expression.

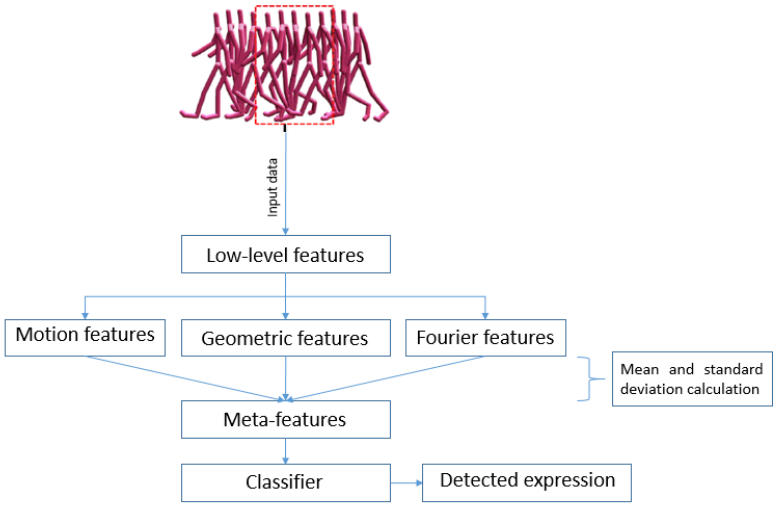


Figure 4.2 : Schéma global de l'approche dite experte de reconnaissance d'expressions : calcul des trois familles de descripteurs bas-niveaux regroupés dans des méta-descripteurs puis classifiés. Source : [Cre+16]

La table 4.1 décrit les 68 descripteurs. Pour chaque descripteur, un méta-descripteur est calculé (moyenne et écart type) pour obtenir un total de 136 valeurs. La figure 4.3 présente les méta-descripteurs pour trois expressions différentes lors d'un mouvement consistant à frapper à une porte. Les relativement bons taux de reconnaissances présentés dans la section 4.5 indiquent que ces descripteurs experts sont globalement discriminants pour les expressions présentes dans les bases de données standards.

Les deux premiers descripteurs proposés cherchent à caractériser la posture du corps de manière globale. Le premier descripteur, V, va calculer le volume global occupé par le squelette pour chaque posture du mouvement en calculant la taille de la boîte englobante du squelette aligné sur les axes du monde. Le second descripteur, θ, est l'angle défini entre la verticale du monde y et l'axe du buste formé à partir du centre du bassin allant au cou. Ce descripteur permet d'avoir l'inclinaison du corps par rapport à la verticale. En effet, en se référant à plusieurs études dans le domaine de la psychologie, une personne à tendance à étendre son corps lors d'expression positive. Au contraire, une personne qui cherche à

Id.	Type de descripteur	Description
V	Espace occupé par le squelette	La taille de la boîte englobante du squelette.
θ	Angle	Les 3 angles induits par le triangle formé entre les deux épaules et le cou. Angle entre la direction verticale du monde y et l'axe formé par le centre du bassin à la tête.
$\mathcal{D}$	Distance	Main droite et le bassin. Main gauche et le bassin. Main droite et épaule droite. Main gauche et épaule gauche. Coude droit et le bassin. Coude gauche et le bassin.
A	Aire	Triangle formé par les deux mains et le cou. Triangle formé par les deux épaules et le cou. Triangle formé par les deux mains et le bassin. Triangle formé par les deux coudes et le bassin.
$\vec{v}$	Vitesse	Mains, Épaules, Bassin, Tête, Coudes.
$\vec{a}$	Accélération	Mains, Épaules, Bassin, Tête, Coudes..
$\mathcal{F}$	Fréquentiel	Transformée de Fourier appliquée au signal de rotation des articulations suivantes : Mains, Épaules, Bassin, Tête, Coudes..

Table 4.1 : Ensemble des descripteurs experts pour reconnaître une expression corporelle.

transmettre une expression négative va adopter une posture plus compacte, notamment en se penchant un peu vers l'avant.

Les distances D entre différentes articulations sont très importantes. Ces distances raffinent la posture en fournissant des informations précises comme le niveau de renfermement ou d'expansion. Par exemple, les distances entre les mains et les épaules, entre les mains et le bassin, entre les coudes et le bassin sont calculées. En plus de ces distances, des descripteurs d'aires A entre certains triplets d'articulations sont ajoutés. Ces triangles fournissent une information sur la forme de la posture. Le lien entre les deux côtés d'un corps est important. Ainsi, les triangles sont formés en prenant une articulation de chaque côté du corps, gauche et droit, et une articulation sur l'axe central du corps, soit le cou, soit le bassin.

De nombreuses études ont montré que certaines articulations précises sont plus importantes pour discriminer différents états émotionnels. Par exemple, Bernhardt *et al.* [BR07] analysent uniquement la vitesse de la main. Nous avons testé et validé l'utilisation d'uniquement les descripteurs suivants : la vitesse absolue $\vec{v}$ et l'accélération absolue $\vec{a}$, des deux mains, des deux épaules, du bassin, de la tête et des deux coudes. La vitesse est la dérivée première de la position de l'articulation cible. L'accélération est la dérivée seconde de cette position.

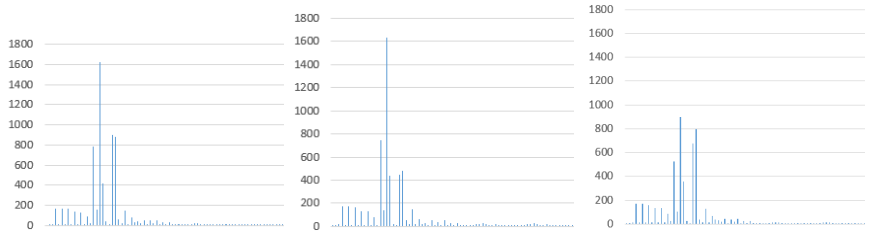


Figure 4.3 : Histogramme des descripteurs calculés lors de l'action de frapper à une porte avec des expressions différentes : à gauche la joie, au milieu la tristesse, à droite la colère. Ces histogrammes montrent que les descripteurs sont discriminants. Pour des raisons de présentations, uniquement les 76 premiers descripteurs du vecteur de 136 descripteurs sont affichés. Source : [Cre+16]

Pour finir, la transformée de Fourier $\mathcal{F}$, fournit une information fréquentielle pour certaines articulations. La transformée de Fourier discrète via l'algorithme de *FFT* sur le signal de rotation de l'articulation donne la phase et l'amplitude du mouvement de l'articulation. Ces valeurs sont concaténées en utilisant chacune des articulations. L'analyse fréquentielle de Fourier est souvent utilisée avec succès dans l'analyse et la synthèse de style avec par exemple Yumer *et al.* [YM16] qui l'utilisent pour transférer le style d'une animation vers une animation neutre.

Ce chapitre présente plusieurs approches, parfois très différentes, de reconnaissance d'expressions dans un geste humain. Les résultats sont présentés dans la section 4.5.

4.2 Mouvement neutre à partir d'un mouvement expressif

Comme présenté dans la section 4.1, les approches basées sur des descripteurs experts fournissent de bons résultats. Cependant, ce type d'approche est limité lorsque l'on travaille sur une grande variété de mouvements. En effet, ces approches fonctionnement mieux lorsque le mouvement est connu et que les descripteurs sont spécialisés pour une action unique. Il serait intéressant d'avoir des approches plus génériques, invariantes au mouvement original et à l'action sous-jacente. Intuitivement, un classifieur fonctionnera mieux si une approche est capable d'extraire une information épurée de l'action sous-jacente réalisée. En s'inspirant du domaine de la synthèse d'animations où les méthodes se basent sur un mouvement neutre afin d'y ajouter un style, nous cherchons à produire ce mouvement neutre pour améliorer la reconnaissance d'expressions corporelles. Il est raisonnable de considérer que le résidu obtenu entre un mouvement expressif et un mouvement neutre contient l'information relative au style. La synthèse automatique d'un mouvement neutre à partir du mouvement expressif n'est pas triviale. Nous proposons de modifier un mouvement par optimisation en formalisant une fonction donnant un score de "neutralité". Même si ce mouvement garde parfois des défauts, il peut servir pour analyser les différences entre le mouvement expressif et le mouvement neutre synthétisé afin d'aider le classifieur à reconnaître l'expression. La synthèse d'un mouvement neutre permet alors d'avoir une méthode automatique de reconnaissance d'expressions invariante au mouvement corporel

exécuté. La figure 4.5 illustre le processus global de génération du mouvement neutre et de l'extraction du résidu pour la classification.

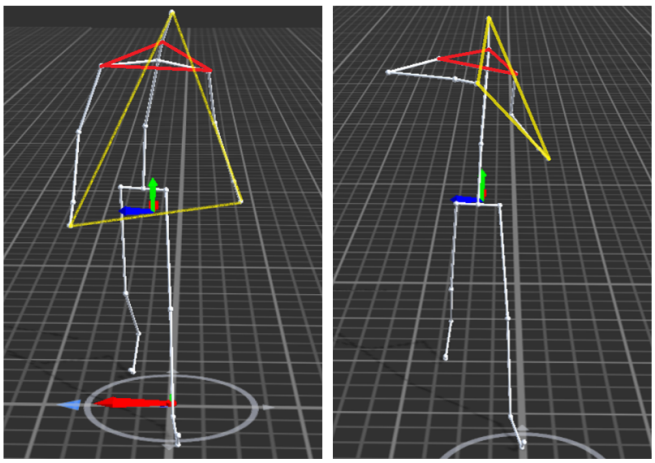


Figure 4.4 : Différentes poses d'un même mouvement avec des expressions corporelles différentes. Cette figure montre les variations de la zone triangulaire formée par les deux épaules et le cou. À gauche, une pose d'une marche déprimée. À droite, une pose d'une marche fière. Source : [Cre+16].

4.2.1 Génération automatique d'un mouvement neutre

Le verrou principal consiste à générer une animation neutre à partir d'un mouvement expressif. La notion de mouvement neutre versus expressif est toujours liée à un contexte. Nous définissons un mouvement neutre comme un mouvement comportant juste une action, c'est-à-dire non porteur d'une marque émotionnelle. Ainsi, une personne décrivant le mouvement ne devrait que qualifier l'action produite et rien d'autre. L'approche se base sur un filtrage des trajectoires de chaque articulation afin de réduire les oscillations dans le mouvement expressif. La seconde étape fait appel à une fonction de coût qui réalise un compromis entre supprimer les "détails" du geste et ne pas trop s'éloigner de l'action initiale. Une optimisation permet plus de contrôle sur le mouvement produit comparé à des méthodes procédurales et demande moins de données qu'une approche à base d'apprentissage.

Dans la thèse d'Arthur Crenn, nous avons proposé [Cre+17b ; Cre+20 ; Cre19] deux manières d'obtenir un mouvement neutre. Les deux approches sont basées sur une optimisation d'une fonction de coût qui décrit le degré de "neutralité" d'une animation. La première fonction de coût se base des caractéristiques cinématiques (distance, vitesse, accélération) pour chaque articulation. L'animation est filtrée par un lissage des trajectoires couplé à un algorithme de cinématique inverse afin de garantir les contraintes anatomiques d'un corps humain. Cette approche permet d'obtenir un mouvement neutre, mais avec une impression très robotique. Pour des raisons de concision, seule la deuxième méthode est présentée ici. Elle produit un mouvement plus humainement réaliste. Elle améliore la définition de la fonction de "neutralité", tout en réalisant l'optimisation de l'animation en la plongeant

dans un espace de dimension réduite représentant l'ensemble des mouvements humains plausibles.

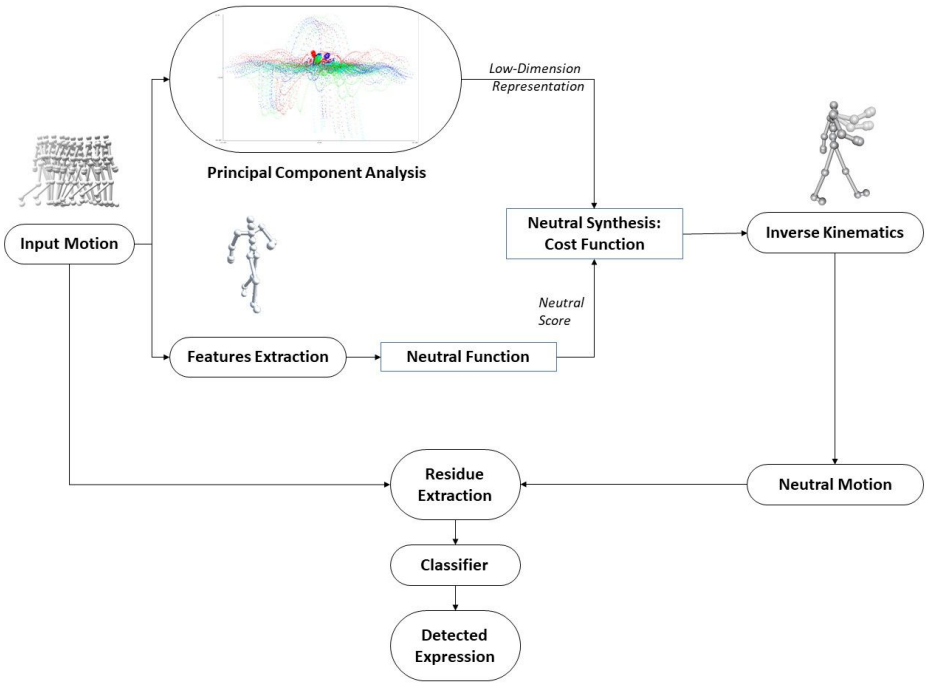


FIGURE 4.5 : À partir du mouvement expressif, nous synthétisons un mouvement neutre grâce à l'optimisation d'une fonction qui donne un score de neutralité pour un mouvement donné. Le résidu entre le mouvement neutre synthétisé et le mouvement original est extrait pour être donné à un classifieur afin de reconnaître l'expression corporelle du mouvement en entrée. Source : [Cre+20]

Mouvement neutre caractérisé par une fonction de coût

Pour transformer automatiquement un mouvement expressif en un mouvement neutre, nous définissons une fonction devant caractériser un mouvement neutre. À partir d'une animation donnée, cette fonction doit calculer une valeur de "neutralité" : plus la valeur est proche de zéro, plus l'animation est censée être neutre. Une optimisation est appliquée pour modifier l'animation en minimisant la fonction de coût. Modifier une animation en optimisant tous les angles ou positions de chaque articulation de chaque pose est trop complexe en termes de dimensions. Par conséquent, l'animation est convertie dans un espace réduit construit par deux étapes d'analyse en composante principale (*ACP*).

La fonction générale caractérisant la "neutralité" d'une animation est décrite dans l'équation 4.1. Elle comporte trois termes. Le premier terme $Neutral$ est un terme de neutralité. Il permet d'évaluer la neutralité d'un mouvement donné. Le deuxième terme $Data$ évalue la distance entre le mouvement neutre synthétisé et le mouvement expressif fourni en entrée du système. Ce terme est utilisé afin que le mouvement neutre réalise la même action que le mouvement original. Le dernier terme est un terme de régularisation $Penalty$ qui ajoute une pénalité si le mouvement neutre synthétisé ne respecte pas les contraintes :

problèmes de fluidité et des pieds qui glissent. Dans ce document, seul le terme $Neutral$, le plus important, est détaillé. Les deux autres termes sont classiques en optimisation et sont décrits en détail dans [Cre19 ; Cre+20].

$$Cost(Motion) = \lambda Neutral(Motion) + \gamma Data(Motion) + \beta Penalty(Motion) \quad (4.1)$$

Afin de pouvoir juger la neutralité d'un mouvement, le terme $Neutral$ se base sur le formalisme de différents descripteurs qualitatifs provenant du domaine de la psychologie [FP18]. Ces descripteurs décrivent qualitativement l'expression d'un mouvement. Ces descripteurs sont séparés en deux parties : les descripteurs de posture et les descripteurs temporels. Le terme $Neutral$ est défini comme une moyenne pondérée décrite dans l'équation 4.2 où α_i correspond au poids du descripteur courant f_i. Le nombre de descripteurs est n.

$$Neutral(Motion) = \frac{\sum_{i=1}^{n} \alpha_i f_i}{\sum_{i=1}^{n} \alpha_i} \quad (4.2)$$

Avant de détailler les différents descripteurs utilisés, les valeurs pertinentes des poids présents dans la fonction de neutralité sont optimisées afin d'obtenir un score tendant vers zéro pour l'ensemble des mouvements neutres extraits des bases de données de mouvements.

Les descripteurs de posture sont inspirés des descripteurs experts présentés dans la section 4.1.

- L'ouverture du corps $f_{bodyOpenness}$ calcule la distance entre les deux bras, entre la main droite et le bassin, entre la main gauche et le bassin et entre les pieds.
- L'inclinaison sagittale du corps $f_{bodyLeaning}$ mesure l'angle entre le vecteur du bassin au cou et l'axe vertical du monde. Ce descripteur donne une indication sur l'inclinaison du corps.
- La droiture du corps $f_{bodyStraightness}$ calcule l'angle de fléchissement de la tête, des genoux et du tronc.

Les descripteurs temporels sont les suivants.

- La puissance du mouvement $f_{mvtPower}$ représente la quantité de force déployée lors du mouvement.
- La fluidité du mouvement $f_{mvtFluidity}$ représente la variation ou non du mouvement global.
- La vitesse du mouvement $f_{mvtSpeed}$ calcule la vitesse à laquelle le mouvement est réalisé.
- La quantité de mouvements des bras $f_{quantityArmsMvt}$ mesure si les bras bougent ou non.
- La régularité du mouvement des bras $f_{regularityArmsMvt}$ cherche à détecter des motifs dans le mouvement des bras.

La fluidité du mouvement est la moyenne de l'accélération pour les deux mains et les deux pieds. La vitesse du mouvement est calculée par la moyenne de la vitesse de chacune des articulations du squelette. La vitesse d'une articulation est simplement définie par la différence de position entre le temps t et le temps $t-1$. La quantité de mouvement des bras est définie comme la distance cumulée couverte par chacune des articulations du bras

lors du mouvement. La transformée de Fourier donne une information sur la régularité du mouvement des bras en sommant les différences d'amplitude du spectre de chacune des articulations des deux bras (épaules, coudes et mains). La transformée de Fourier est un formalisme classique pour extraire différents composants depuis une animation [LG15].

Représentation d'un mouvement dans un espace de dimension réduite

Le but de ce changement d'espace est de pouvoir transposer un mouvement dans une représentation compacte, mais précise. En effet, l'espace original d'un mouvement est de trop grande dimension pour pouvoir être utilisé directement dans un processus d'optimisation. Afin de résoudre ce problème, nous proposons d'utiliser deux analyses en composantes principales (*ACP*) imbriquées : l'une travaillant sur les poses et l'autre travaillant sur l'aspect temporel. L'*ACP* cherche à réduire les dimensions d'un espace en construisant un ensemble de vecteurs pertinents. La combinaison linéaire de ces vecteurs permet de retrouver l'espace initial. Il s'agit d'obtenir le résumé le plus pertinent de nos données originales. Pour cela, la matrice des variances-covariances permet de réaliser cette réduction de dimension en analysant la dispersion des données initiales. L'*ACP* est classique en analyse des données et en animations [Jol11 ; Gaf+12]. Il aurait sûrement été possible de construire un espace réduit à partir d'un auto-encodeur comme présenté dans la section 5.3, mais l'*ACP* propose une représentation plus compacte. En effet, un auto-encodeur construit un espace latent très représentatif, mais pas forcément avec une dimension réduite [Hol+15]. Comme une optimisation est réalisée sur les paramètres, la taille de l'espace est importante et l'approche à base d'*ACP* a cet avantage.

Afin d'aboutir à une représentation compacte d'un mouvement, deux phases d'*ACP* sont réalisées : une première sur toutes les postures du corps sans se soucier du temps, une deuxième sur un mouvement entier à partir des valeurs de la première. La figure 4.6 présente le cheminement de ce processus.

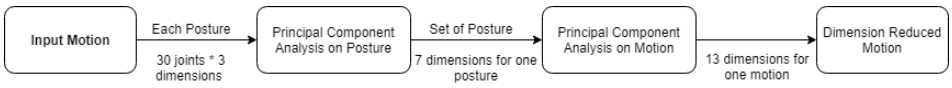


Figure 4.6 : Processus en deux phases d'*ACP* pour la réduction de dimension d'un mouvement. Source : [Cre+20]

Optimisation

Afin de modifier l'animation dans l'espace latent modélisé par les deux *ACP*, la fonction de coût est optimisée par la méthode de la stratégie d'évolution *CMA-ES* (*Covariance Matrix Adaptation Evolution Strategies*). Cette méthode d'optimisation a pour avantage d'être une méthode ne nécessitant pas de dérivées pour des problèmes d'optimisation non linéaire ou non convexe. Il est probable que d'autres algorithmes d'optimisation pourraient résoudre le problème. *CMA-ES* est utilisé classiquement dans le domaine de l'optimisation de mouvements physiques avec de très bons résultats [Won+17].

4.2.2 Classification du résidu entre mouvement neutre et expressif

Le mouvement original expressif est modifié vers un mouvement neutre en minimisant la fonction de coût présentée précédemment. La classification utilise la différence entre les deux mouvements : le résidu. Il est calculé dans le domaine fréquentiel sur chaque degré de liberté des articulations entre le mouvement expressif et le mouvement neutre. Cette idée que le domaine fréquentiel est un bon outil pour traiter du style est présente par exemple dans les travaux de Yumer et Mitra [YM16]. Dans la représentation fréquentielle, la magnitude contient principalement l'information de mouvement ainsi que l'expression pour une animation donnée. Le résidu entre l'animation neutre et l'animation expressive est calculé avec la différence de magnitude pour chaque degré de liberté de chaque articulation. Les détails des calculs se trouvent dans [Cre19 ; Cre+20]. La section 4.5 présente l'ensemble des résultats de toutes les approches présentées.

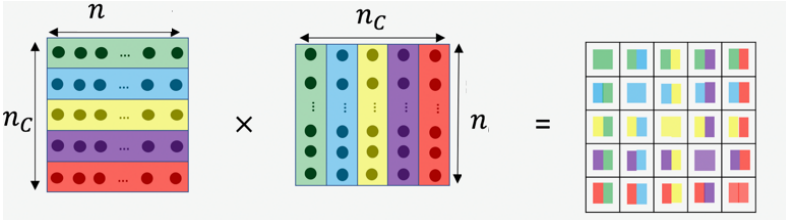


Figure 4.7 : La matrice de Gram se calcule à partir de n_c caractéristique de taille n qui forment une matrice. Cette matrice est multipliée par sa transposée pour donner chaque couple de corrélation dans la matrice de Gram. Cette matrice met en valeur les corrélations entre les caractéristiques qui sont souvent plus importantes que les caractéristiques elles-mêmes.

4.3 Espace latent et matrice de Gram

Dans le domaine du transfert de style entre images, Gatys *et al.* [GEB16] remarquent que les couches d'un réseau de convolution apprennent des propriétés de plus en plus haut niveau au fur et à mesure que l'on avance dans les couches. En faisant usage d'un réseau de neurones de l'état de l'art dans la reconnaissance d'images, *VGG16* [SZ14], ils parviennent à séparer dans une certaine mesure le style et le contenu d'une image. Cette idée que chaque couche d'un réseau produit des descripteurs peut être transposée aux animations. Un espace latent dédié aux animations permet d'en extraire des descripteurs. Pour mesurer la corrélation entre descripteurs, il est possible de calculer la matrice de Gram. Cette matrice est une indication pertinente de style pour les images : deux images ayant une matrice de Gram similaire ont un style similaire. Ce point devrait se vérifier également pour les animations et donc aider pour la reconnaissance de style. Nous avons mené une étude afin de nous en assurer tout en mesurant l'importance de l'espace latent et de la matrice de Gram en comparaison à des descripteurs classiques. Puis la meilleure configuration de cette section est comparée aux autres approches dans la section 4.5.

4.3.1 Matrice de Gram

La matrice de Gram d'un vecteur de propriétés X est définie par : $G(X) = X \times X^t$. Elle peut s'interpréter comme la corrélation entre les différentes propriétés du vecteur X. Cette matrice permet de mettre en valeur les corrélations entre les vecteurs de propriété, ce qui est souvent plus important que les propriétés elles-mêmes. Si une entrée dans la matrice de Gram a une valeur proche de 0, cela signifie que les deux propriétés ne s'activent pas simultanément (non-corrélation). Réciproquement, si une entrée a une grande valeur, cela signifie que les 2 propriétés s'activent simultanément (corrélation). Par exemple, avec n_c propriétés de taille n, la figure 4.7 présente le calcul de matrice de Gram avec chaque couple de corrélation.

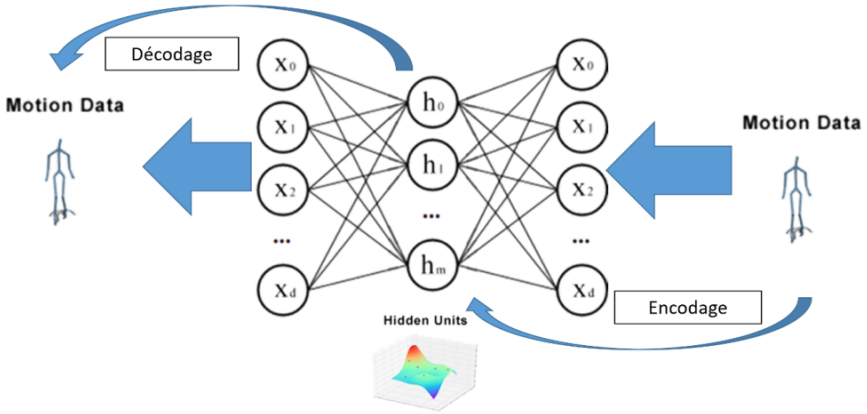


Figure 4.8 : Un auto-encodeur crée un espace latent représentant les mouvements humains. Source : [HSK16].

4.3.2 Auto-encodeur

Un auto-encodeur (*AE*) est un réseau neuronal qui tente de reconstruire son entrée [Hol+15]. Les *AE* sont structurés pour prendre une entrée, la transformer en une représentation différente appelée le code, dans un espace latent. À partir de ce code, la deuxième partie vise à reconstruire l'entrée originale. Les couches de l'*AE* qui créent le code sont appelées l'encodeur et les couches qui essaient de reconstruire l'original à partir du code sont appelées le décodeur. L'objectif d'un *AE* est que le code (l'espace latent) représente de manière plus efficace que les données brutes les variations. L'espace latent cherche un code exhaustif, uniformément réparti, continu, etc. Par exemple, n'importe quelle valeur de code donne une sortie plausible. Deux entrées proches donnent deux codes proches. Les *AE* servent par exemple [TBL18] à corriger des données aberrantes, à améliorer la classification, à améliorer les interpolations, à aider au calcul de similarité, etc.

Holden *et al.* [HSK16] ont proposé un *AE* pour créer un espace latent représentant les mouvements humains. L'encodeur reçoit en entrée 73 valeurs : 73 valeurs par pose et 240 poses. Le squelette comporte 22 articulations. Les 73 valeurs d'une pose se décomposent comme ceci : 22 coordonnées dans l'espace pour les articulations, 3 déplacements élémentaires, et 4 données booléennes de contact au sol. L'encodeur applique trois convolutions

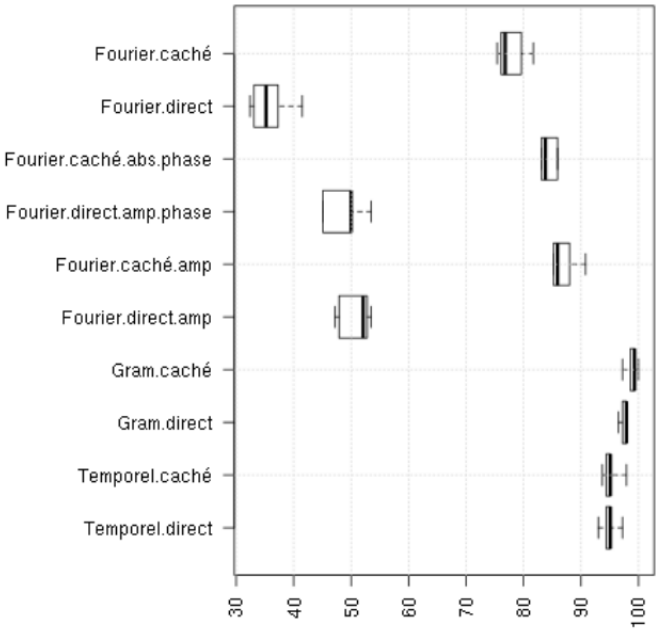


Figure 4.9 : Résultats de classification avec différentes transformations des données. Abscisse, taux de réussite du classificateur avec la transformation choisie. Les boîtes à moustache donnent la répartition des résultats des 5 répétitions de l'expérience. "temporel" : données sans traitement spécifique. "Gram" : matrice de Gram. "Fourier" : transformée de Fourier forme algébrique. "amp" : transformée de Fourier, amplitude. "amp phase" : transformée de Fourier, amplitude et phase. Un *AE* est ajouté ou non à ces 5 types de données : "direct" signifie sans *AE* ; "caché" signifie avec *AE*.

1D successives couplées à des *max-pooling* pour réduire la dimension. La représentation du mouvement dans l'espace latent a une taille de 256. Le décodeur est symétrique et applique des convolutions et des *depooling*. Les données retrouvent leur dimension de départ. L'*AE* présenté ici est modifié de deux manières. La première consiste à représenter les coordonnées des articulations dans un repère relatif à l'origine du membre : l'épaule pour le bras, la hanche pour les jambes, etc. L'idée intuitive est de rendre chaque partie du corps plus indépendante du reste du corps. La deuxième amélioration ajoute un terme supplémentaire à la fonction de coût afin de rendre le code plus épars (*sparse*) [San+17].

4.3.3 Étude comparative

Il est proposé ici qu'un *AE* couplé à la matrice de Gram est efficace pour la classification de l'expression portée par un geste humain et ceci pour n'importe quelle action réalisée par la personne. Pour vérifier cette supposition, l'approche est comparée aux différentes transformations suivantes :

- identité (pas de transformation) ;
- matrice de Gram ;
- transformée de Fourier, forme algébrique (le classifieur reçoit en entrée $R(F[X])$ et $I(F[X])$) ;

— transformée de Fourier, forme exponentielle (le classifieur reçoit en entrée |F[X]| et arg(F[X])) ;
— transformée de Fourier, amplitude seule (le classifieur reçoit en entrée seulement |F[X]|) .

Chacune de ces transformations s'applique soit aux données brutes (les quaternions de chaque articulation), soit aux codes produits par l'*AE* décrit précédemment. La série de valeurs obtenue est donnée au classifieur. Le classifieur est un réseau de neurones complètement connecté (un perceptron multicouche). L'étude est menée sur la base de données de mouvements *SIGGRAPH-DB* [Xia+15a]. Ce n'est pas la base de données la plus exigeante, car les mouvements proviennent d'animateurs et les styles sont souvent exagérés. Mais cette étude permet de savoir si l'*AE* et Gram sont réellement pertinents. Le processus de validation suit le schéma classique de validation croisée *k-fold* : répétition du partage avec 75% des données pour l'apprentissage et 25% pour le test. Pour assurer que l'apprentissage ne soit pas déséquilibré, c'est-à-dire, qu'il n'y ait pas de style sous-représenté, 75% des échantillons de chacun des huit styles sont récupérés pour constituer le jeu de données d'apprentissage. Chaque ensemble de données entraînement/test est mélangé cinq fois. Les résultats donnés sur la figure 4.9 montrent que la meilleure configuration de cette étude est d'utiliser le code produit par l'*AE* mis en valeur par la matrice de Gram. Cette approche est comparée à l'état de l'art et à nos autres approches dans la section 4.5. Ces travaux n'ont pas été publiés, mais ils disposent d'un potentiel certain. Pour surpasser réellement les taux de l'état de l'art actuel, il faudrait revisiter l'*AE* en explorant le potentiel des *Variational-AE* [Hab+17 ; LKC21 ; Liu+20] ou des réseaux de diffusion [Ram+21a].

4.4 Méta-analyse quantitative des descripteurs d'expressions

Pour l'analyse, la reconnaissance ou la synthèse d'expressions dans un mouvement, le principal défi consiste à trouver une liste de paramètres pertinents capables de "coder" n'importe quelle expression indépendamment de l'action ou de la personne réalisant le geste. Les travaux dans les domaines de la psychologie, des arts, de la vision par ordinateur et de la synthèse d'images ont proposé de nombreux paramètres, appelés souvent des descripteurs. Pour la reconnaissance d'expressions, le critère d'explicabilité peut être secondaire. Cependant, pour l'analyse et la synthèse, disposer de descripteurs explicables et compréhensibles par toutes les personnes travaillant sur l'expressivité dans les mouvements est un but important. En effet, pour de nombreuses applications, ces paramètres sont étudiés, partagés, modifiés par un humain. L'expert veut comprendre pourquoi telle ou telle expression est présente, et souvent il aimerait pouvoir la modifier. Par exemple, un animateur voudrait ajouter, exagérer ou transférer des expressions sur tout type de geste ou sur tout type de personnage virtuel. Un psychologue veut trouver un lien entre certains paramètres pour mieux comprendre les émotions.

Les poids d'un réseau de neurones, tel que celui présenté dans la section 4.3 restent inaccessibles à un humain et les descripteurs calculés sont difficiles à formaliser. D'un autre côté, les paramètres experts présentés dans la section 4.1 sont souvent intuitifs. Par exemple, certains paramètres ont une intuition géométrique comme l'inclinaison de la

colonne vertébrale, l'ouverture des épaules ou l'angle des pieds, etc. D'autres paramètres ont une intuition physique comme une mesure de l'énergie de chaque articulation ou une mesure de "douceur" de la trajectoire. Une personne penchée vers l'avant ou avec un mouvement à faible énergie reflétera un sentiment plus négatif qu'une personne au torse bombé ou avec un mouvement dynamique. Les descripteurs experts présentés jusque là sont moins performants en termes de classification que l'*AE*/Gram ou que la génération d'un geste neutre (voir la section résultats en 4.5), ce qui laisse penser que leur capacité de généralisation à des variations externes serait moindre.

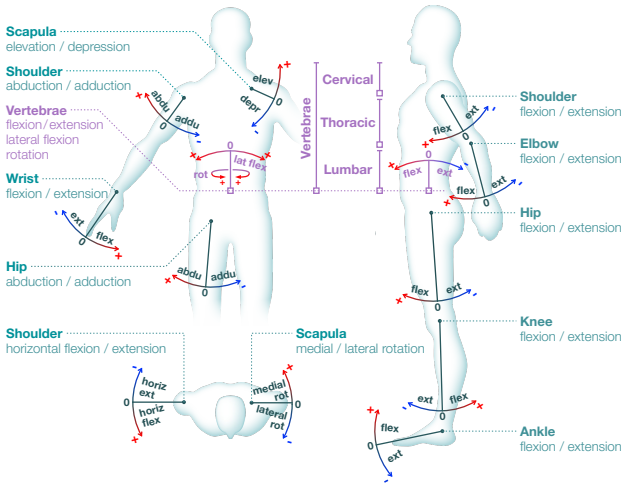
Nous avons voulu pousser plus loin l'étude de l'état de l'art afin de mieux définir et catégoriser les descripteurs experts tout en leur associant des valeurs quantitatives. Si des plages de valeurs de chaque descripteur aident à améliorer la reconnaissance, ils pourront également servir en animation. Cependant, les valeurs numériques dans les publications sont hétérogènes dans leur définition, dans leur mesure et dans leur unité. Le travail consiste à faire une revue systématique de tous les paramètres décrits dans les différents domaines de recherche et de les unifier en une liste quantifiée de paramètres expressifs. Pour valider les valeurs numériques des paramètres, ils sont également mesurés sur la base de données *Emilya* [FP14]. La comparaison des valeurs issues de la méta-analyse et de la base de données de mouvements devrait montrer une corrélation. La publication [Mah+22] donne de nombreux détails supplémentaires.

4.4.1 Processus de la méta-analyse

Dans la section 4.1, une première version de ces descripteurs est proposée. Elle est largement améliorée ici, en précisant clairement des familles de descripteurs, ainsi qu'en leur associant des valeurs numériques. Ces améliorations sont possibles grâce à une méta-analyse plus avancée de la littérature. Cette méta-analyse porte sur la qualification et la quantification des descripteurs compréhensibles responsables de l'expression des émotions dans le corps et le mouvement humains. Elle est basée sur méthode PRISMA [Moh+09]. Pour la première étape, trois bases de données de publications ont été interrogées pour extraire les publications dites "état de l'art". Treize états de l'art ont été identifiés avec au moins une section traitant de l'expression des émotions dans le mouvement et/ou la posture humaine, quel que soit le domaine d'application (reconnaissance, synthèse, applications médicales, etc.). Pour la deuxième étape, 1228 références sont extraites de ces états de l'art pour finalement garder 216 articles traitant des émotions dans le mouvement et la posture d'un corps humain. L'inclusion est basée sur les critères suivants : l'article fournit au moins un descripteur extractible du mouvement et/ou de la posture du corps humain ; au moins un des descripteurs peut être facilement compris ; l'article fournit des mesures pour au moins un descripteur. Nous avons exclu les résultats portant sur les humeurs (phénomènes de plus longue durée), la douleur (qui peut être considérée comme un phénomène émotionnel distinct), et plus généralement les phénomènes affectifs liés au domaine de la santé.

4.4.2 Les caractéristiques

Chaque descripteur présenté ci-dessous est composé d'un label, d'une description, d'une liste des publications le référençant, et d'une plage de valeurs numériques (une moyenne



Name*	References
ankle flexion	[GCF12]
cervical vert. flex.	[GCF12][Wal98][Jam32b][Roe+09b][Mic+09][Bus+07]
cervical vert. lat. flex.	[Wal98][Bus+07]
cervical vert. rot.	[Bus+07]
elbow flexion	[GCF12][OG07a][Roe+09b]
hip flexion	[GCF12][OG07a][Roe+09b]
knee flexion	[GCF12][OG07a][Roe+09b]
scapula elevation	[GCF12][Wal98]
scapula medial rot.	[OG07a][Wal98][KBS11]
shoulder abduction	[KBS11]
shoulder flex.	[GCF12][OG07b][Roe+09b]
shoulder horiz. flex.	[KBS11]
thoracic vert. flex.	[GCF12]
vert. flex.	[GCF12][KBS11][Jam32b][Roe+09b][MGC87]
vert. lat. flex.	[GCF12]
vert. rot.	[GCF12]
wrist flexion	[GCF12]

*Notes :
vert. stands for *vertebrae*,
flex. stands for *flexion*,
lat. stands for *lateral* ,
rot. stands for *rotation*,
horiz. stands for *horizontal*

FIGURE 4.10 : Les rotations des descripteurs biomécaniques issues de la littérature sont unifiées. Source : [Mah+22].

et un écart type). La plage numérique correspond à deux pôles opposés. Une telle représentation bipolaire permet de déterminer le niveau d'influence d'un descripteur expressif pour une émotion donnée, par rapport à un état neutre, qui est déterminé statistiquement. Les descripteurs sont agrégés depuis plusieurs articles, les définitions et les valeurs numériques sont uniformisées. L'ensemble des descripteurs sont regroupés en trois familles, allant des descripteurs de plus bas niveau, très factuels sur le mouvement, vers des descripteurs plus élaborés faisant appel à des concepts plus complexes.

Les caractéristiques biomécaniques désignent l'ensemble des descripteurs expressifs de bas niveau. Ce sont typiquement les angles des articulations comme la flexion des cervicales, l'abduction de l'épaule, l'inclinaison de la colonne, etc. La colonne "Références" dans la table de la figure 4.10 en donne la liste. Bien que ces descripteurs semblent les plus simples, l'hétérogénéité des descriptions entraîne une grande confusion, et parfois des expériences difficilement reproductibles. La description du squelette de référence est parfois succincte. La description textuelle peut prêter à confusion. La représentation numérique des angles (Euler, rotation autour d'un axe, etc.) demande de la minutie dans les conversions. Afin de fusionner ces descripteurs, nous avons utilisé les connaissances du domaine de l'anatomie fonctionnelle [Kap18a ; Kap18b ; Kap19] pour regrouper toutes les rotations articulaires référencées dans la littérature. La figure 4.10 fournit une illustration de ces rotations, ainsi que les publications associées.

Les caractéristiques de Laban sont les descripteurs liés aux travaux du chorégraphe du vingtième siècle, Rudolf Laban (voir la section 2.3.1). La seconde partie de la table 4.2 présente une liste des principales composantes des facteurs de Laban relatifs à l'effort et à la forme ainsi que les études qui les mesurent. Ces facteurs sont regroupés en deux catégories distinctes :

— l'effort pour les facteurs dynamiques (comment le mouvement progresse dans une séquence temporelle) ;

— la forme pour les facteurs posturaux (comment la forme du corps change).

L'effort est subdivisé en composantes d'espace, de temps, de poids et de flux. La forme peut être liée à l'effort en l'exprimant avec les mêmes composantes. Laban n'a pas associé ses descriptifs à un formalisme scientifique précis. De nombreux articles proposent d'en calculer les termes, mais des divergences existent. Notre article [Mah+22] propose une unification autour de la manière de calculer ces facteurs qui n'est pas détaillée ici.

Les caractéristiques de haut niveau regroupent les caractéristiques plus élaborées, dont l'explication repose sur des adjectifs descriptifs faisant appel à une certaine compréhension des propriétés générales d'un mouvement ou de la démarche. Contrairement aux caractéristiques biomécaniques qui sont destinées à décrire les mouvements du corps de manière très factuelle, ou aux caractéristiques de Laban qui sont souvent géométriques ou physiques, les caractéristiques de haut niveau permettent de décrire les mouvements et les postures d'une manière plus conceptuelle. Elles peuvent sembler subjectives, mais elles peuvent en fait être définies comme une combinaison de facteurs inférieurs. De plus, ces caractéristiques sont évolutives, dans le sens où elles peuvent être utilisées pour décrire les mouvements d'un membre aussi bien que d'un corps entier. Ces caractéristiques sont néanmoins toujours des valeurs numériques. La première partie du tableau 4.2 en donne une liste associée aux articles.

Validation

La méta-analyse agrège des données de domaines variés, ce qui peut sembler risqué. Pour valider leur valeur, ces mêmes descripteurs sont calculés directement sur une base de données de mouvements annotée émotionnellement, dans le but de valider les résultats de la méta-analyse. De nombreuses bases de données sont disponibles dans le domaine de l'analyse du mouvement (voir la section 2.3.3). Nous avons choisi la base de données *Emilya* [FP14], car elle fournit un grand nombre de fichiers d'animation avec une diversité d'actions et d'expressions. Kleinsmith *et al.* [KB13] mettent en garde contre l'influence que peuvent avoir les actions spontanées ou actées sur la pertinence des analyses. Cependant, la plupart des articles de la méta-analyse se base des mouvements actés donc le biais est déjà présent dans les valeurs numériques de la méta-analyse. De plus, nous pensons que la gamme de valeurs obtenues peut-être particulièrement utile pour les applications de synthèse d'animations où les émotions jouées, souvent plus exagérées et stéréotypées, ne semblent pas être un problème et peuvent même être une qualité requise. La comparaison entre les résultats de la méta-analyse et de l'analyse d'*Emilya* est présenté sur la figure 4.11. Toutes les valeurs dans le graphique sont normalisées.

4.4.3 Valeurs numériques et limitations

La table 4.12 donne les plages de valeurs pour les descripteurs de la méta-analyse. Leur grande diversité révèle la complexité de la question des expressions corporelles. Il y a des entrées manquantes dans la table, mais elles pourraient sûrement être complétées par des valeurs calculées sur la base de mouvements, puisque l'expérience montre qu'il y a une corrélation entre les deux. Pour les valeurs présentes, la comparaison numérique entre les valeurs issues des publications et la base *Emilya* permet de valider les travaux des scientifiques du domaine : ce qui est plutôt une bonne nouvelle. Des travaux sont en cours

Table 4.2 : Caractéristiques de haut niveau issues de la méta-analyse. Source : [Mah+22].

Features based on the effort and shape factors of Laban				
Name	Description	Negative pole	Positive pole	References
laban flow effort	Degree of control, resistance or introversion during movement progression	Bound, tense movement progression, feeling of control.	Free, loose movement progression, feeling of relaxation.	[GCF12][Mon+99] [GCF10][Bro+97]
laban flow shape	Amount of self-centeredness in body posture configuration	Closing, shrinking, contracting, orienting inward	Opening, growing, expanding, orienting outward	[GCF12][GCF10] [Jam32b][Pia+13]
laban space effort	Level of expansion or contraction during movement performance	Direct, straight line direction, tapered movement progression, feeling of narrowness	Indirect, undulating line direction, flexible movement progression, feeling of large space	[GCF12][Wal98] [Mon+99][SSI03] [Mei89][GP06][WS86] [GCF10][Bro+97] [Pia+13][MKI09] [NAK02]
laban space shape	Occupation of limb on transversal plane	Enclosing, narrowing, decreasing transversal occupation	Spreading, widening, increasing transversal occupation	[GCF12][FP18] [GCF10][MKI09] [NAK02][Mei89]
laban time effort	Time impression during movement performance	Sudden, high speed, short time period, ephemeral feeling	Sustained, low speed, long time period, sensation of eternity	[GCF12][Mon+99] [SSI03][DF07][Mei89] [GP06][WS86][Pol+01] [FP18][GCF10] [MKI09][Sam+13] [NAK02][Bus+07] [Zac+13a]
laban time shape	Orientation of body bosture on lateral plane	Retreating, hallowing, orienting backward	Advancing, bulging, orienting forward	[Mei89][FP18][Pia+13] [MKI09][NAK02]
laban weight effort	Quantity of force, energy and gravity perceived during movement performance	Strong, sensation of weight, high gravity resistance, energy liberation	Light, feeling of weightlessness, low gravity resitance	[GCF12][Wal98] [Mon+99][SSI03] [Mei89][GP06][WS86] [FP18][GCF10] [Bro+97][Pia+13] [MKI09][Sam+13] [NAK02][MGC87]
laban weight shape	Configuration of body posture on vertical axis	Sinking, shortening, orienting downard	Rising, lengthening, orienting upward	[Wal98][Mei89][FP18] [MKI09][NAK02] [Jam32b]
Features based on gait properties, descriptive adjectives and general movement properties				
Name	Description	Negative pole	Positive pole	References
approach	Overwhole body orientation toward an object or situation	Avoiding, turning away	Approching, turning toward	[Jam32b][AL14]
exaggeration	Amplification of timing/speed/amplitude of movement	Unexagerrated (tends to neutral movement)	Highly exagerrated and stereotyped	[Bro+97]
movement activity	Overall impression of movement quantity, considering both amplitude and frequency	Tends to immobility	A lot of movements	[Wal98][Mon+99] [WS86][Pia+13] [Glo+11][FP18]
movement amp.	Spatial and/or angular range of movement primitive	Small amplitude	Large amplitude	[DF07][GCF10]
movement anticipation	Preparation time before action execution. Anticipation usually consists in movement in the opposite direction of the main action	Lower anticipation time	Higher anticipation time	[GCF10]
movement exertion	Needed time for main action progression. Exertion is easily recognizable in a jump movement as the airborne phase	Lower exertion time	Higher exertion time	[GCF10]
movement settle	Stabilization phase right after exertion. Preceeds rest pose or new movement execution	Lower settle time	Higher settle time.	[GCF10]
regularity	Amount of changes in phrasing, tempo and speed during overall performance	Irregular	Regular	[DF07][FP18][Pia+13]
smoothness	General impression of fluency during movement performance	Jerky and with a lot of stop and go	Smooth/fluent movements	[Mon+99][DF07] [FP18][Bro+97] [Pia+13][Glo+11] [AL14]
walk arm swing	Amplitude of front/back arms sway during walk	Higher swing amplitude	Lower swing	[Mic+09][MGC87]
walk lat. body sway	Lateral sway of upper body during walk, left and right movements are undifferentiated	Reduced sway, impression of rigid walk	Exagerated sway, a feeling of "drunk" walk	[Mic+09]
walk speed	Translationnal velocity	Lower velocity	Higher velocity	[GCF12][KG16] [Mic+09]
walk step frequency	Number of footstep during a specific time period	Lower frequency	Higher frequency	[GCF12][KG16]
walk stride duration	Duration between two consecutive footsteps	Lower duration	Higher duration	[KG16]
walk stride length	Distance between two consecutive footsteps	Lower stride length	Higher stride length	[GCF12][KG16] [MGC87]
walk vertical head sway	Up-Down head sway during walk	Lower vertical sway	Higher vertical sway	[Mic+09]

pour évaluer la pertinence de ces descripteurs pour de la reconnaissance d'émotion. Les résultats préliminaires sont présentés dans la section suivante. De plus, l'objectif d'utiliser ces plages de valeurs pour aider à la synthèse de gestes expressifs est présenté dans le chapitre 5.

Pour aller plus loin dans la compréhension des expressions dans les gestes, il y a sûrement d'autres descripteurs pertinents et intuitifs à inventer. Même les descripteurs de haut niveau peinent à caractériser les expressions subtiles. La méta-analyse montre que les émotions de base sont relativement bien couvertes. Cependant, le vocabulaire des émotions est riche et la manière dont une émotion s'exprime et transparaît dans les gestes peut être très mystérieuse. Même si certaines expressions peuvent être un mélange d'expressions basiques, les schémas numériques ont besoin de valeurs, de moyens de calculer ce "mélange".

4.5 Comparaisons des approches et conclusion

Nous avons évalué les différentes méthodes sur quatre bases de données présentées dans la section 2.3.3. Trois de ces bases ont été réalisées par capture de mouvements d'acteurs jouant diverses actions comme marcher, frapper, porter, lancer, etc. La dernière base, nommée *SIGGRAPH-DB*, est utilisée en synthèse d'animations et a été produite par des animateurs. Elle contient des expressions un peu plus exagérées et caricaturales que les autres. La base *Emilya* comporte un grand nombre d'animations avec une variété d'actions, d'expressions pour une dizaine d'acteurs, afin de tester la généricité des méthodes. Nous l'utilisons souvent comme référence pour les comparaisons.

Table 4.3 : Comparaison de nos méthodes avec l'état de l'art.

Autres publications	**UCLIC**	**MPI**	**SIG-DB**	**Emilya**	**Famille**	**Explicable**
Par un humain [Nes15]				41%		
[Nes15]				75%	Desc. Exp.	++
[DAS17]	87.30%				Laban+LSTM(DL)	+
[PBD22]				82.3%	Autoenc.+Graph(DL)	-
[Bey+21]				91.3%	Conv. (DL)	-
[Shi+22]				92%	GraphNN (DL)	-
Nos approches						
Descripteurs experts [Cre+16]	78%		93%		Desc. Exp.	++
Mouvement neutre 1 [Cre+17b]		50%	50%		Hybride	+
Mouvement neutre 2 [Cre+20]	74%	78.6%	98.8%	82.2%	Hybride	+
Espace Latent et Gram (non publié)			98%	84%	Autoenc.(DL)	-
Méta-analyse (travaux en cours)					Desc. Exp.	++

La table 4.3 présente la comparaison du taux de reconnaissance de nos méthodes avec les méthodes de l'état de l'art dont les résultats sont disponibles sur une de ces quatre bases. En effet, de nombreuses publications ne sont pas dans ce tableau, car elles ne proposent des résultats qu'avec des bases de données propriétaires non disponibles [TBZ16 ; AMD19 ; Liu+21 ; Nor+18], ce qui est dommage. Néanmoins, ce tableau permet de voir l'évolution des approches en général. Trois critères sont donnés.

— Taux de bonnes classifications.

— Famille de méthode : Descripteurs experts, *Deep Learning* (DL) en précisant le type de réseaux, hybride.
— Explicable [AB18] : allant de - à ++. Ce critère subjectif essaie d'indiquer si l'approche est explicable et compréhensible afin de pouvoir servir à la synthèse d'animations relativement facilement.

On constate que les descripteurs experts sont progressivement remplacés par des approches exploitant l'apprentissage profond au détriment souvent de l'explicabilité. Les réseaux pourraient produire des caractéristiques compréhensibles, mais en visant la performance de reconnaissance en priorité, on ne modélise pas le réseau avec cet objectif. Nos deux travaux générant un mouvement neutre sont hybrides : plus complexes que les approches basées sur des descripteurs experts, mais également un peu moins explicables et donc modifiables. La deuxième approche modélise un espace latent à base d'*ACP* et se rapproche donc des approches d'apprentissage profond. Dans ces résultats, nous confirmons également l'idée que plus le mouvement neutre synthétisé est réaliste, meilleur est le taux de reconnaissance.

Nous avons également remarqué que les publications ne présentent jamais de validation croisée sur les sujets : apprendre avec les gestes d'une sous-partie des sujets et tester sur les sujets restants. Des recherches que nous sommes sur le point de soumettre suggèrent que l'évaluation d'un modèle d'apprentissage sur des individus inconnus diminue drastiquement les taux de reconnaissance. En comparaison, le domaine de la reconnaissance d'actions fait déjà une validation croisée en séparant les sujets [WHK20 ; Qin+22]. Ceci donne un bon critère sur la capacité de généralisation de la méthode. Pour les expressions, il est difficile d'obtenir les chiffres de l'état de l'art sur ce point, mais il faudrait que la communauté intègre ce critère.

Publications liées

- Arthur Crenn, Rizwan Ahmed Khan, Alexandre Meyer et Saida Bouakaz. "Body expression recognition from animated 3D skeleton". In : (2016), p. 1-7
- Arthur Crenn, Alexandre Meyer, Rizwan Ahmed Khan, Hubert Konik et Saïda Bouakaz. "Toward an Efficient Body Expression Recognition Based on the Synthesis of a Neutral Movement". In : ICMI 2017 Proceedings of the 19th ACM International Conference on Multimodal Interaction (2017)
- Arthur Crenn, Alexandre Meyer, Hubert Konik, Rizwan Ahmed Khan et Saida Bouakaz. "Generic body expression recognition based on synthesis of realistic neutral motion". In : *IEEE Access* 8 (2020), p. 207758-207767
- Mehdi-Antoine Mahfoudi, Alexandre Meyer, Thibaut Gaudin, Axel Buendia et Saida Bouakaz. "Emotion Expression in Human Body Posture and Movement : A Survey on Intelligible Motion Factors, Quantification and Validation". In : *IEEE Transactions on Affective Computing* (2022), p. 1-24

Thèses

- Arthur Crenn. "Capture et transfert d'expression de visages d'enfants pour l'interaction avec des mondes virtuels". Theses. Université Claude Bernard Lyon 1, LIRIS, 2019

- Mehdi-Antoine Mahfoudi. “Génération procédurale d’animations porteuses d’expressions : approche temps réel et interactive”. Theses. Université Claude Bernard Lyon 1, LIRIS, 2023

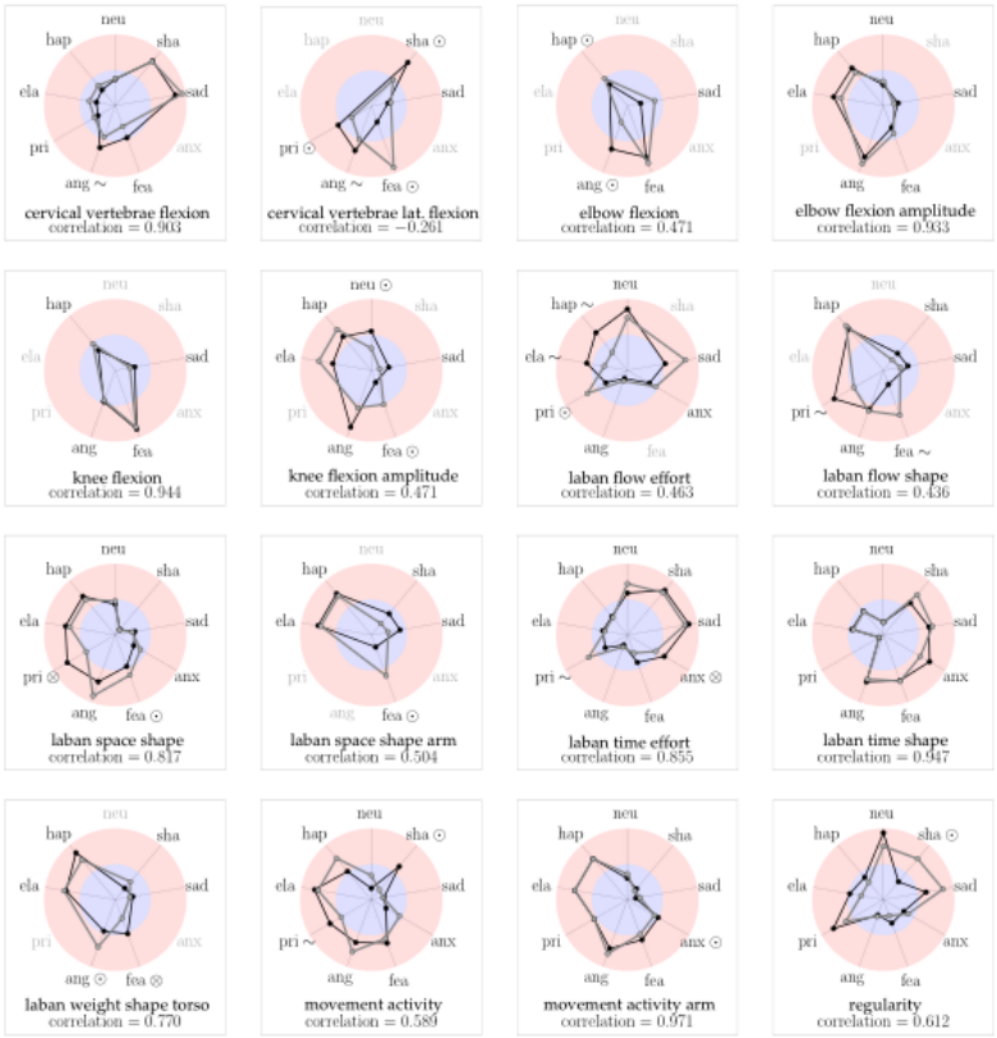
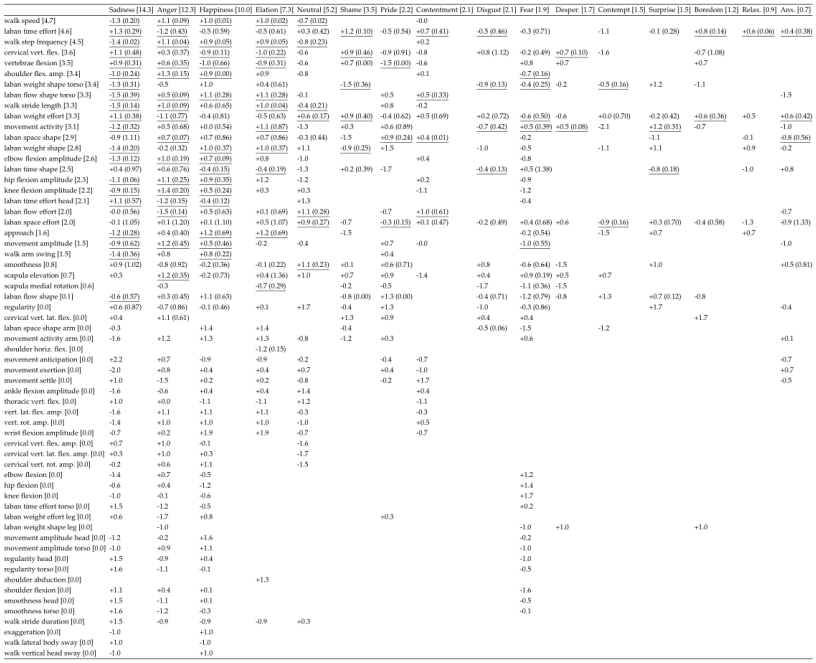


FIGURE 4.11 : Comparaison de 16 valeurs normalisées des descripteurs entre la méta-analyse et l'étude sur la base de mouvements *Emilya* [FP14]. Les 15 autres descripteurs sont présentés dans [Mah+22]. Chaque émotion dont les valeurs ont des signes opposés est suivie de l'un des trois symboles suivants : ∼ indique une cellule de la méta-analyse qui n'est pas perceptible, ⊙ indique une cellule de la méta-analyse ne contenant que la valeur d'une seule expérience et ⊗ pour tous les autres cas. Source : [Mah+22].



	Sadness [14.3]	Anger [12.3]	Happiness [10.0]	Elation [7.3]	Neutral [5.2]	Shame [3.5]	Pride [2.2]	Contentment [2.1]	Disgust [2.1]	Fear [1.9]	Desper. [1.7]	Contempt [1.5]	Surprise [1.5]	Boredom [1.2]	Relax. [0.9]	Anx. [0.7]
walk speed [4.7]	-1.3 (0.20)	+1.1 (0.09)	+1.0 (0.01)	+1.0 (0.02)	-0.7 (0.02)			-0.0								
laban time effort [4.6]	+1.3 (0.29)	-1.2 (0.43)	-0.5 (0.59)	-0.5 (0.61)	+0.3 (0.42)	+1.2 (0.10)	-0.5 (0.54)	+0.7 (0.41)	-0.5 (0.46)	-0.3 (0.71)		-1.1	-0.1 (0.28)	+0.8 (0.14)	+0.6 (0.06)	+0.4 (0.38)
walk step frequency [4.5]	-1.4 (0.02)	+1.1 (0.04)	+0.9 (0.05)	+0.9 (0.05)	-0.8 (0.23)			+0.2								
cervical vert. flex. [3.6]	+1.1 (0.48)	+0.3 (0.37)	-0.9 (0.11)	-1.0 (0.22)	-0.6	+0.9 (0.46)	-0.9 (0.91)	-0.8	+0.8 (1.12)	-0.2 (0.49)	+0.7 (0.10)	-1.6		-0.7 (1.08)		
vertebrae flexion [3.5]	+0.9 (0.31)	+0.6 (0.35)	-1.0 (0.66)	-0.9 (0.31)	-0.6	+0.7 (0.00)	-1.5 (0.00)	-0.6		+0.8	+0.7			+0.7		
shoulder flex. amp. [3.4]	-1.0 (0.24)	+1.3 (0.15)	+0.9 (0.00)	+0.9	-0.8			+0.1		-0.7 (0.16)						
laban weight shape torso [3.4]	-1.3 (0.31)	-0.5	+1.0	+0.4 (0.61)		-1.5 (0.36)			-0.9 (0.13)	-0.4 (0.25)	-0.2	-0.5 (0.16)	+1.2	-1.1		
laban flow shape torso [3.3]	-1.5 (0.39)	+0.5 (0.09)	+1.1 (0.28)	+1.1 (0.28)	-0.1		+0.5	+0.5 (0.33)								-1.5
walk stride length [3.3]	-1.5 (0.14)	+1.0 (0.09)	+0.6 (0.65)	+1.0 (0.04)	-0.4 (0.21)		+0.8	-0.2								
laban weight effort [3.3]	+1.1 (0.38)	-1.1 (0.77)	-0.4 (0.81)	-0.5 (0.63)	+0.6 (0.17)	+0.9 (0.40)	-0.4 (0.62)	+0.5 (0.69)	+0.2 (0.72)	-0.6 (0.50)	-0.6	+0.0 (0.70)	-0.2 (0.42)	+0.6 (0.36)	+0.5	+0.6 (0.42)
movement activity [3.1]	-1.2 (0.32)	+0.5 (0.68)	+0.0 (0.54)	+1.1 (0.87)	-1.3	+0.3	+0.6 (0.89)		-0.7 (0.42)	+0.5 (0.39)	+0.5 (0.08)	-2.1	+1.2 (0.31)	-0.7		-1.0
laban space shape [2.9]	-0.9 (1.11)	+0.7 (0.07)	+0.7 (0.86)	+0.7 (0.86)	-0.3 (0.44)	-1.5	+0.9 (0.24)	+0.4 (0.01)		-0.2			-1.1		-0.1	-0.8 (0.56)
laban weight shape [2.8]	-1.4 (0.20)	-0.2 (0.32)	+1.0 (0.37)	+1.0 (0.37)	+1.1	-0.9 (0.25)	+1.5		-1.0	-0.5		-1.1	+1.1		+0.9	-0.2
elbow flexion amplitude [2.6]	-1.3 (0.12)	+1.0 (0.19)	+0.7 (0.09)	+0.8	-1.0			+0.4		-0.8						
laban time shape [2.5]	+0.4 (0.97)	+0.6 (0.76)	-0.4 (0.15)	-0.4 (0.19)	-1.3	+0.2 (0.39)	-1.7		-0.4 (0.13)	+0.5 (1.38)			-0.8 (0.18)		-1.0	+0.8
hip flexion amplitude [2.3]	-1.1 (0.06)	+1.1 (0.25)	+0.9 (0.35)	+1.2	-1.2			+0.2		-0.9						
knee flexion amplitude [2.2]	-0.9 (0.15)	+1.4 (0.20)	+0.5 (0.24)	+0.3	+0.3			-1.1		-1.2						
laban time effort head [2.1]	+1.1 (0.57)	-1.2 (0.15)	-0.4 (0.12)		+1.3					-0.4						
laban flow effort [2.0]	-0.0 (0.56)	-1.5 (0.14)	+0.5 (0.63)	+0.1 (0.69)	+1.1 (0.28)		-0.7	+1.0 (0.61)								-0.7
laban space effort [2.0]	-0.1 (1.05)	+0.1 (1.20)	+0.1 (1.10)	+0.5 (1.07)	+0.9 (0.27)	-0.7	-0.3 (0.15)	+0.1 (0.47)	-0.2 (0.49)	+0.4 (0.68)	+0.6	-0.9 (0.16)	+0.3 (0.70)	-0.4 (0.58)	-1.3	-0.9 (1.33)
approach [1.6]	-1.2 (0.28)	+0.4 (0.40)	+1.2 (0.69)	+1.2 (0.69)		-1.5				-0.2 (0.54)		-1.5	+0.7		+0.7	
movement amplitude [1.5]	-0.9 (0.62)	+1.2 (0.45)	+0.5 (0.46)	-0.2	-0.4		+0.7	-0.0		-1.0 (0.55)						-1.0
walk arm swing [1.5]	-1.4 (0.36)	+0.8	+0.8 (0.22)				+0.4									
smoothness [0.8]	+0.9 (1.02)	-0.8 (0.92)	-0.2 (0.36)	-0.1 (0.22)	+1.1 (0.23)	+0.1	+0.6 (0.71)		+0.8	-0.6 (0.64)	-1.5		+1.0			+0.5 (0.81)
scapula elevation [0.7]	+0.3	+1.2 (0.35)	-0.2 (0.73)	+0.4 (1.36)	+1.0	+0.7	+0.9	-1.4	+0.4	+0.9 (0.19)	+0.5	+0.7				
scapula medial rotation [0.6]		-0.3		-0.7 (0.29)		-0.2	-0.5		-1.7	-1.1 (0.36)	-1.5					
laban flow shape [0.1]	-0.6 (0.57)	+0.3 (0.45)	+1.1 (0.63)			-0.8 (0.00)	+1.3 (0.00)		-0.4 (0.71)	-1.2 (0.79)	-0.8	+1.3	+0.7 (0.12)	-0.8		
regularity [0.0]	+0.6 (0.87)	-0.7 (0.86)	-0.1 (0.46)	+0.1	+1.7	-0.4	+1.3		-1.0	-0.3 (0.86)			+1.7			-0.4
cervical vert. lat. flex. [0.0]	+0.4	+1.1 (0.61)				+1.3	+0.9		+0.4	+0.4				+1.7		
laban space shape arm [0.0]	-0.3		+1.4	+1.4		-0.4			-0.5 (0.06)	-1.5		-1.2				
movement activity arm [0.0]	-1.6	+1.2	+1.3	+1.3	-0.8	-1.2	+0.3			+0.6						+0.1
shoulder horiz. flex. [0.0]				-1.2 (0.15)												
movement anticipation [0.0]	+2.2	+0.7	-0.9	-0.9	-0.2		-0.4	-0.7								-0.7
movement exertion [0.0]	-2.0	+0.8	+0.4	+0.4	+0.7		+0.4	-1.0								+0.7
movement settle [0.0]	+1.0	-1.5	+0.2	+0.2	-0.8		-0.2	+1.7								-0.5
ankle flexion amplitude [0.0]	-1.6	-0.6	+0.4	+0.4	+1.4			+0.4								
thoracic vert. flex. [0.0]	+1.0	+0.0	-1.1	-1.1	+1.2			-1.1								
vert. lat. flex. amp. [0.0]	-1.6	+1.1	+1.1	+1.1	-0.3			-0.3								
vert. rot. amp. [0.0]	-1.4	+1.0	+1.0	+1.0	-1.0			+0.5								
wrist flexion amplitude [0.0]	-0.7	+0.2	+1.9	+1.9	-0.7			-0.7								
cervical vert. flex. amp. [0.0]	+0.7	+1.0	-0.1		-1.6											
cervical vert. lat. flex. amp. [0.0]	+0.3	+1.0	+0.3		-1.7											
cervical vert. rot. amp. [0.0]	-0.2	+0.6	+1.1		-1.5											
elbow flexion [0.0]	-1.4	+0.7	-0.5							+1.2						
hip flexion [0.0]	-0.6	+0.4	-1.2							+1.4						
knee flexion [0.0]	-1.0	-0.1	-0.6							+1.7						
laban time effort torso [0.0]	+1.5	-1.2	-0.5							+0.2						
laban weight effort leg [0.0]	+0.6	-1.7	+0.8				+0.3									
laban weight shape leg [0.0]		-1.0								-1.0	+1.0			+1.0		
movement amplitude head [0.0]	-1.2	-0.2	+1.6							-0.2						
movement amplitude torso [0.0]	-1.0	+0.9	+1.1							-1.0						
regularity head [0.0]	+1.5	-0.9	+0.4							-1.0						
regularity torso [0.0]	+1.6	-1.1	-0.1							-0.5						
shoulder abduction [0.0]				+1.3												
shoulder flexion [0.0]	+1.1	+0.4	+0.1							-1.6						
smoothness head [0.0]	+1.5	-1.1	+0.1							-0.5						
smoothness torso [0.0]	+1.6	-1.2	-0.3							-0.1						
walk stride duration [0.0]	+1.5	-0.9	-0.9	-0.9	+0.3											
exaggeration [0.0]	-1.0		+1.0													
walk lateral body sway [0.0]	+1.0		-1.0													
walk vertical head sway [0.0]	-1.0		+1.0													

FIGURE 4.12 : Résultats de la méta-analyse : dans chaque cellule, la première valeur correspond à la moyenne et la valeur entre parenthèses à l'écart-type. L'absence de parenthèse indique que la cellule est le résultat d'une seule valeur. Les cellules remarquables sont soulignées ($\mathrm{unnoticeable}(\mathbf{X})$ strictement inférieur à un). Une cellule est vide si aucune étude n'a fourni de valeur numérique. Chaque descripteur et chaque étiquette d'émotion est suivi d'un nombre entre crochets qui représente le niveau de perceptibilité. Source : [Mah+22].

5 Animation : contrôle, édition et styles

Table des matières du chapitre

Les deux chapitres précédents sont consacrés à la reconnaissance d'expressions émises par les mouvements du visage et du corps, souvent en s'inspirant de techniques de synthèse d'animations. Par exemple, dans le cas du visage, nous nous sommes basés sur l'analyse de la texture du visage, élément qui, dans ce chapitre, est capturé depuis une caméra et transformé pour permettre la synthèse de rides de peaux lors de l'animation de visages. Pour la reconnaissance d'expressions corporelles, nous avons proposé une approche utilisant des outils de synthèse d'animations pour générer automatiquement un mouvement neutre (l'action) à partir d'une animation portant une expression. Nous avons également réalisé une méta-analyse de toutes les publications relatives aux expressions corporelles avec pour objectif de répertorier tous les descripteurs de style associés à des valeurs numériques. Dans ce chapitre, ces descripteurs sont également considérés pour réaliser de la synthèse de styles dans une animation. Les travaux présentés ont pour objectif de synthétiser ou de modifier une animation afin de pouvoir prendre en compte un style comme illustré sur la figure 5.1. Nous préférons le terme "style" ici au terme "expression", car en synthèse d'animations les qualificatifs d'animations sont plus variés que les expressions avec par exemple les styles âgé, fatigué, zombi, enfantin, sexy, etc. Une grande majorité de nos travaux concernent l'humain en mouvement, car ils représentent un centre d'intérêt important pour les applications. Nous présentons néanmoins des travaux de génération procédurale de locomotion de créatures à n pattes (fourmis, araignées, robot, etc.) où la démarche peut être interprétée comme un style particulier, mais également comme des travaux préliminaires visant à proposer un modèle procédural et général (créatures et humains) d'animation avec possibilité d'éditer le style.

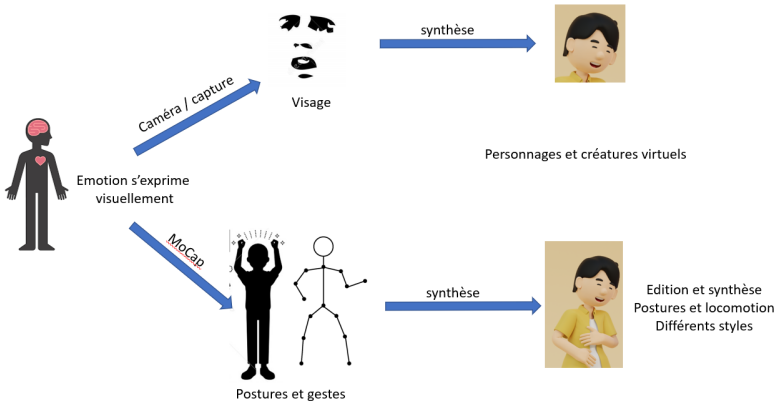


FIGURE 5.1 : Dans ce chapitre, nous abordons des techniques de synthèse d'animations faciale en nous basant sur de la capture simple, une approche de locomotion générique de créatures à n pattes (insectes, araignées, robots, etc.), ainsi que des techniques d'édition de postures avec un début de prise en compte de descripteurs expressifs.

Plus précisément, les domaines de l'animation faciale, de l'animation procédurale de créatures, de l'édition de poses et de l'animation procédurale de style sont abordés. Nous nous sommes particulièrement concentrés sur l'accessibilité des solutions proposées. Cette accessibilité se traduit par une facilité d'utilisation, même pour un utilisateur novice, et par des besoins limités en matériel et en données. Certaines méthodes comme les approches procédurales ne nécessitent aucune donnée. D'autres nécessitent des données, comme l'animation faciale et l'édition de poses par réseaux de neurones, mais nous avons cherché

à ce que ces données soient simples à obtenir et que l'entraînement soit raisonnable. Le lien avec l'analyse et la vision par ordinateur est toujours présent : les textures ou les mouvements sont capturés par des approches de vision par ordinateur, les descripteurs que nous cherchons à éditer sont issus d'une méta-analyse des publications en psychologie, en vision et en IHM. Nos travaux montrent qu'il existe un lien entre ces différentes communautés.

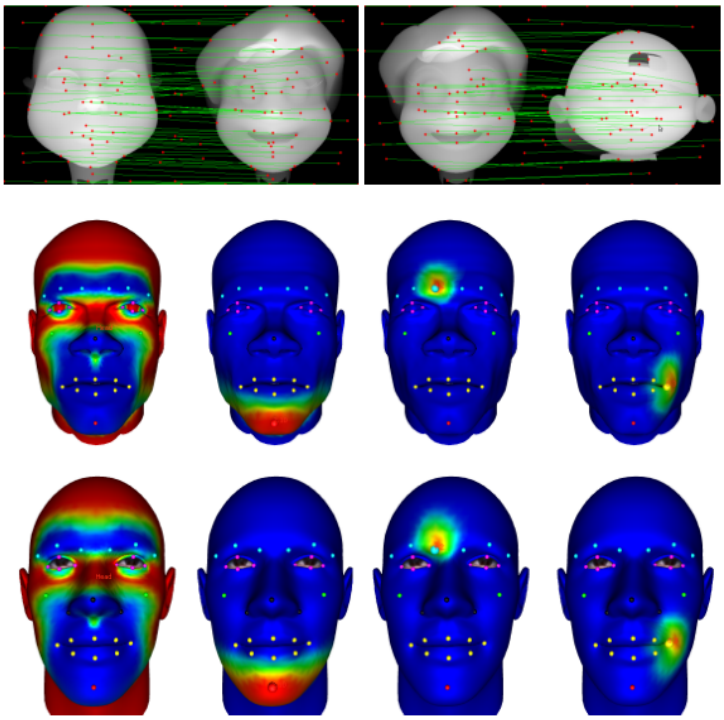


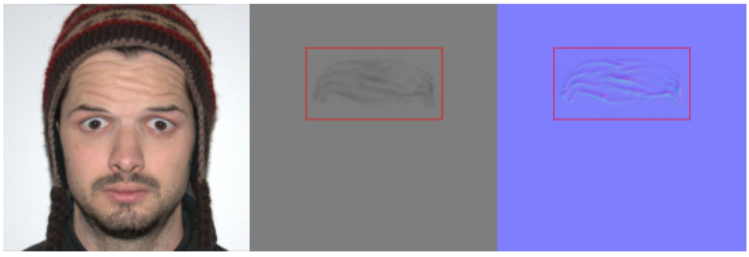
Figure 5.2 : Résultat du transfert automatique des poids du *skinning*. En haut, des points caractéristiques sont trouvés puis appariés automatiquement. Puis, une mise en correspondance dense permet de transférer les influences du *skinning*. Le visage du milieu est le visage source, le visage cible (en bas) est paramétrisé automatiquement par transfert.

5.1 Capture et transfert des éléments du visage donnant de l'expressivité

Les travaux de cette section concernent la problématique de l'animation faciale pour des applications en temps réel alliant capture, transfert et animation. Ce sont les travaux les plus anciens, certains choix se feraient sûrement différemment aujourd'hui. Ils sont présentés dans un but d'exhaustivité. De nombreux compléments peuvent se retrouver dans la thèse de Ludovic Dutreve [Dut11].

L'animation faciale est l'un des points clés dans le réalisme des scènes 3D qui mettent en scène des personnages virtuels. Ceci s'explique principalement par les raisons suivantes : les nombreux muscles qui composent le visage permettent de générer une multitude d'expressions ; ensuite, notre faculté de perception permet de détecter et d'analyser les

mouvements les plus fins. De ce fait, il est très difficile de créer une animation de qualité sans un travail manuel long et fastidieux. De plus, dans ces travaux, l'objectif du temps réel et de l'accessibilité est primordial. Trois problèmes principaux sont abordés. Le premier concerne la paramétrisation du visage pour l'animation. La paramétrisation a pour but de définir des moyens de contrôle pour pouvoir déformer et animer le visage. Le second s'oriente sur l'animation, et plus particulièrement sur le transfert d'animation. Le but est de proposer une méthode qui permet d'animer le visage d'un personnage à partir de données variées. Ces données peuvent être issues d'un système de capture de mouvements, ou bien elles peuvent être obtenues à partir de l'animation d'un personnage virtuel qui existe déjà. Enfin, nous nous sommes concentrés sur les détails fins liés à l'animation comme les rides. Bien que ces rides soient souvent fines et discrètes, elles jouent un rôle important dans la perception et l'analyse des émotions [CBM09]. Une technique d'acquisition monoculaire est proposée et complétée par une méthode à base de poses références pour synthétiser dynamiquement les détails fins d'animation sur le visage. Ces travaux restent utilisables dans de nombreuses applications aujourd'hui, car ils se basent sur des techniques connues et largement répandues dans des applications interactives, à savoir le *skinning* [Kav+07] pour la paramétrisation du visage et les cartes de normales pour un rendu par pixel (en opposition au rendu par sommet).



(a) expression déformée (b) image ratio (c) carte de normales

Figure 5.3 : À partir de l'image capturée comportant les déformations (a), nous calculons une image de ratio comportant l'inclinaison du gradient (b) puis la carte de normales (c). La normale correspond à la direction perpendiculaire de la surface et permet de calculer un éclairage fin.

5.1.1 Paramétrisation du visage par transfert

Pour animer un maillage de visage, il doit être paramétrisé (*rigging* [SS18]). Nous ciblons l'animation par *skinning* [Kav+07] car c'est l'approche la plus répandue dans l'industrie. Pour le *skinning* des points de contrôles (os) doivent être définis et chaque sommet du maillage doit être associé aux os qui l'influencent. Pour trouver la paramétrisation d'un nouveau visage, nous sommes partis du principe qu'il est plus judicieux de réutiliser une paramétrisation existante plutôt que d'en recréer une nouvelle à partir de rien. Nous avons ainsi développé une nouvelle méthode de transfert de paramétrisation en nous basant sur la mise en correspondance de surface. Cette technique permet de détecter des points repères dans chacun des deux visages 3D. Les paires de points repères obtenues permettent ensuite de transférer les points de contrôle du visage source au visage cible via une mise en correspondance de surfaces denses pour transférer les influences du *skinning*. Cette approche de transfert de paramétrisation est entièrement automatique, mais l'intervention

de l'utilisateur reste néanmoins possible s'il souhaite personnaliser les résultats. La figure 5.2 montre des exemples de résultats de transfert des paramètres de *skinning* d'un maillage à un autre.

Figure 5.4 : La première ligne est le visage source en position neutre et avec six expressions. Les trois autres lignes montrent les visages cibles sur lesquels les expressions ont été transférées automatiquement.

5.1.2 Capture et transfert

Pour l'animation en temps réel, nous proposons une méthode de capture et de transfert d'un mouvement de visage réel depuis une webcam vers un modèle 3D de visage. L'information capturée est double : les mouvements des points de contrôle et une carte de normales enrichissant le visage cible de ridules et de plis de peau (voir la figure 5.3). Le transfert des mouvements de points de contrôle se base sur une interpolation par *RBF* (*Radial Basis Function*) [Buh00]. Comme l'animation se base sur les points de contrôle du *skinning*, elle permet avec un même processus d'animer un visage 3D à partir d'une animation existante d'un autre visage 3D à la morphologie différente et d'animer un visage 3D à partir d'un système de capture de mouvements 2D (ou 3D). Le système offre une indépendance entre les régions qui permet de n'effectuer qu'un transfert partiel de l'animation ou de fusionner plusieurs sources dans une seule animation. La capture des points de contrôle peut se faire dorénavant sans marqueur et en temps réel grâce aux approches basées sur l'apprentissage profond [Gao+21]. Le transfert proposé ici est une interpolation par *RBF*, il est probable que les outils récents d'apprentissage profond permettraient d'obtenir des résultats plus réalistes en prenant en compte les déformations non linéaires. L'avantage de l'interpolation par *RBF* est qu'elle ne nécessite pas de long temps d'apprentissage, ni de quantité de données importante. Un maillage 3D de visage suffit.

Afin de générer les détails de plis de peau sur un visage en temps réel, le système proposé se base sur un ensemble restreint de poses de référence alliant des paires de déformations fines et grandes échelles. Les déformations à grandes échelles sont les positions des points

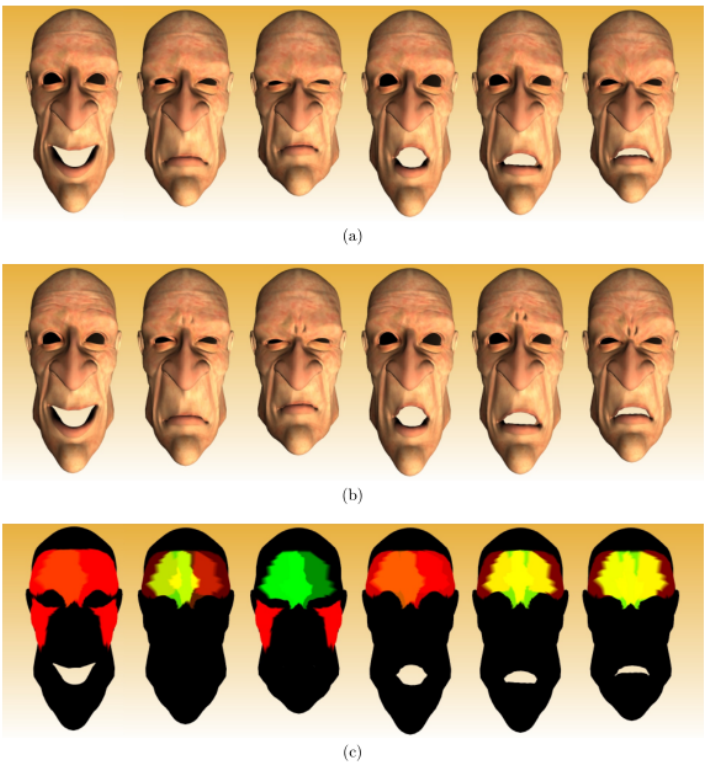


Figure 5.5 : La partie (a) montre les 6 expressions de base (joie, tristesse, dégoût, surprise, peur, colère) sans détail. La partie (b) montre les mêmes expressions en utilisant notre méthode de rides dynamiques avec nos données capturées. La partie (c) montre l'influence des poses de références sur le maillage. Les pixels rouges et verts correspondent à deux cartes de normales différentes, les pixels jaunes sont influencés par les deux cartes de normales.

de contrôle, les os du *skinning*. Les déformations fines sont une carte de normales. Cet ensemble de paires utilisé dans un processus d'interpolation permet de générer à la volée une nouvelle carte de normales pour d'autres positions des points de contrôle. L'utilisation de carte de normales est largement utilisée pour le rendu de scène 3D. Pour obtenir ces cartes de normales, nous avons développé une technique d'acquisition qui permet d'extraire les détails d'animation d'une personne réelle à partir d'un unique appareil de capture (caméra ou webcam). L'acquisition se base sur une approche de *Shape from Shading (SfS)* illustré sur la figure 5.3 qui retrouve la normale en chaque pixel en inversant l'équation de calcul de rendu diffus [MBB07]. L'inversion de l'équation donne l'inclinaison, mais l'orientation reste ambiguë : le problème du *SfS* est sous contraint et mal posé [PF06]. Les ambiguïtés d'orientation restantes sont résolues grâce à la connaissance a priori des lignes de direction des rides sur un visage. La couleur intrinsèque de la peau est obtenue par la moyenne du voisinage de l'image capturée. Cette technique ne nécessite qu'un environnement faiblement contrôlé et peu de moyens matériels. L'utilisation d'une couche de détails fins qui se rajoute à la couche d'animation large offre une flexibilité importante. Son implémentation ne présente pas de difficultés techniques et ne modifie pas les chaînes d'animation et de rendu habituelles. Il est facile de l'activer ou de la désactiver dans le

cadre de l'utilisation de niveaux de détails. Bien sûr, l'ambiguïté d'orientation de la normale pourrait être résolue aujourd'hui par apprentissage [Zha+22], mais finalement dans cette approche un expert a repéré la régularité de direction des rides de peaux de visage et cela est suffisant. La figure 5.5 montre des exemples des six expressions de base avec et sans les cartes de rides et la figure 5.4 montre des transferts d'expression de visage.

5.1.3 Conclusion

Les travaux d'animations faciales présentées ici portent sur quatre points : une approche de transfert de paramétrisations (*rigging* par *skinning*) d'un maillage de visage comportant les paramètres vers un maillage nu ; une approche de transfert de mouvements depuis un visage réel capturé par une caméra ; une approche de capture des rides de peau à partir d'une unique caméra et une approche de synthèse de rides de peau en fonction des mouvements du visage. Ces approches sont toutes temps réel, ne nécessitent que très peu d'interventions manuelles et utilisent une unique webcam. Elles permettent d'animer un maillage de visage de manière expressive, en capturant les expressions et mouvements avec la caméra dans un environnement non complexe (un bureau classique suffit). Elles sont à mettre en relation avec les approches de reconnaissance d'expressions faciales de la section 3.1 qui utilisent également les textures de visages pour reconnaître les expressions. Les approches récentes se basent sur des réseaux, mais elles cherchent toujours à utiliser la capture de détails fins associés aux mouvements des points de contrôles [Cao+21 ; BB18]. Les animations du corps sont un domaine connexe comportant un fort potentiel de travaux à réaliser comme le montre la suite de ce chapitre.

Figure 5.6 : La génération procédurale permet d'animer en temps réel des créatures à n pattes évoluant dans un environnement dynamique. Source : [Kar+11]

5.2 Animation procédurale

Les approches d'animations de personnages virtuels sont très souvent basées sur des données issues de capture de mouvements. Cependant, pour de nombreux animaux réels

ou imaginaires, les données de mouvements sont difficilement captables comme pour les quadrupèdes, les arachnides, les insectes, les reptiles, les robots ou toutes sortes de créatures imaginaires. La solution est alors de se tourner vers les approches procédurales qui disposent de nombreux avantages comme d'être extrêmement contrôlable, compréhensible par un animateur, facile à modifier, très rapide à calculer et sans avoir besoin de données. Un de nos objectifs à long terme est de proposer un système d'animation procédurale générique à tout type de personnages virtuels (humains, animaux, robot, créatures imaginaires). Cette section se focalise en premier sur toutes les créatures terrestres à n pattes en laissant un peu de côté les humains. Ces travaux ont été réalisés dans le cadre d'une collaboration avec l'entreprise *Spirops*. Une suite est présentée dans la section 5.4.2 plus centrée sur les humains virtuels.

La tâche la plus commune que les créatures virtuelles effectuent est la locomotion. La figure 5.6 illustre le type de mondes virtuels ciblés. La richesse de ces animations est notamment due à la variété des morphologies animées (par exemple des chiens, araignées, fourmis, lézards, etc.) et à la variété de leurs démarches (par exemple galop et trot pour les chevaux). Savoir gérer cette diversité est l'un des premiers défis lors de la création d'un système générique réutilisable capable d'animer ces différentes créatures à n pattes. De plus, ces créatures à n pattes naviguent dans des environnements complexes (terrains, escaliers, etc.), avec différents types d'obstacles (stable, instable, dynamique, etc.). Pour cela et afin de générer une simulation crédible, le système d'animation doit s'adapter à l'environnement et produire des animations de locomotion avec des franchissements d'obstacles, des traversées des terrains irréguliers, des évitements d'objets dynamiques, etc.

Figure 5.7 : Cycle de locomotion : les barres représentent la phase de vol des quatre pattes. L'utilisateur édite la barre verte pour changer la synchronisation avec les autres pattes et le temps de vol. Source : [Kar+11]

La plupart des personnages terrestres se déplacent d'un endroit à un autre en mettant un pied devant l'autre de manière répétitive jusqu'à atteindre le point d'intérêt (la cible). Lors du mouvement normal d'un pied, deux phases principales peuvent être distinguées : phase au sol et phase de vol. Pendant la phase d'appui, le pied reste bloqué au sol. Sinon, le pied vole selon une courbe parabolique sans aucun contact avec le sol. Un cycle de locomotion (démarche) est l'acte de répéter ces mouvements des pieds selon un certain rythme ou tempo comme montré sur la figure 5.7.

5.2.1 Contrôleur à trois niveaux

Le système est composé de trois modules présentés sur la figure 5.8 à gauche.

— Contrôle de la créature : le module central qui gère la locomotion globale de la créature à n pattes. Elle repose sur les deux autres modules pour calculer le mouvement

des pieds et le déplacement du corps. Ce module s'occupe également du mouvement de la colonne de bassins quand le corps en dispose (lézard, chien, etc.).

— Gestion de la démarche : ce module impose le rythme du mouvement des pieds selon le tempo défini par l'utilisateur.

— Planification de la trajectoire des pieds : ce module évalue pour chaque pied toutes les cibles et trajectoires possibles en temps réel et choisit parmi toutes les possibilités le meilleur couple (trajectoire, cible), c'est-à-dire la meilleure trajectoire 3D qui traverse l'environnement vers la meilleure cible (empreinte du pied). Ce module s'appuie sur un algorithme de construction de trajectoires en 3D basée sur une représentation discrète de l'environnement mise à jour en prenant en compte les objets dynamiques.

Un système de cinématique inverse (*Inverse Kinematics - IK*) permet de calculer la position des articulations intermédiaires des jambes. En animant le bassin, les pieds et les jambes avec ces trois modules et l'*IK*, une animation complète de la créature est générée. Dans ces travaux, les créatures ciblées ont toutes uniquement des pattes reliées au corps, elles ne disposent pas de bras comme un humain.

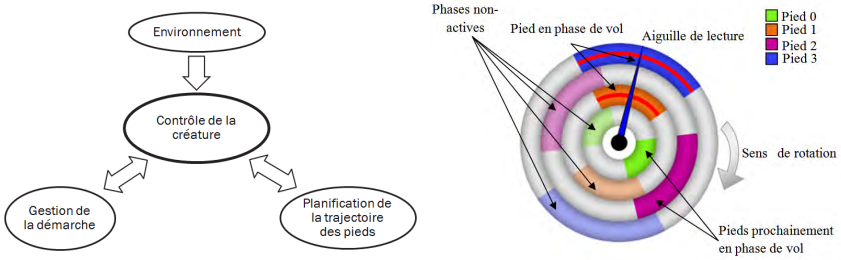


Figure 5.8 : À gauche : l'approche d'animation procédurale se compose de trois modules qui récupèrent des informations de l'environnement. À droite : la locomotion cyclique est naturellement bien représentée par un schéma circulaire avec un cercle pour chaque pied. Source : [Kar+11]

5.2.2 Contrôle de la créature et démarche

Le module de contrôle de la créature gère deux tâches. La première est la gestion du mouvement des pieds où à chaque pas de simulation et selon la démarche, certains pieds vont entrer en phase de vol. Pour chacun de ces pieds, ce module calcule une empreinte préférée selon la vitesse de la créature et l'environnement. Ensuite, la planification est appelée pour produire le meilleur couple : trajectoire et cible du pied. Les pieds au sol sont bloqués à leur position, les glissements ne sont pas autorisés. La deuxième tâche est le calcul du mouvement 3D du bassin qui est détaillé en 5.2.4. Le mouvement 2D du bassin sur le plan horizontal est calculé en utilisant la vitesse et l'orientation. L'élévation et l'inclinaison du bassin sont calculées à partir du sol sous la créature.

Le module de gestion de la démarche organise et visualise le tempo des pieds et effectue les transitions entre différents styles de mouvements. Comme la locomotion est cyclique, la représentation circulaire semble naturelle, comme illustrée sur la figure 5.8 à droite. Chaque cercle représente un pied. Sa phase de vol est représentée par le secteur coloré.

L'aiguille active et désactive les secteurs en fonction de sa position actuelle. L'activation d'un secteur signifie que le pied correspondant entre dans sa phase de vol.

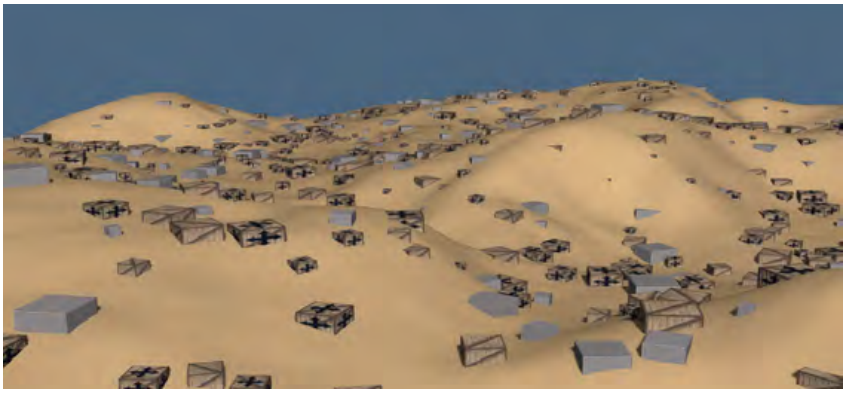


Figure 5.9 : Un exemple d'environnement complexe et dynamique où les créatures doivent pouvoir se déplacer. Source : [Kar+11]

5.2.3 Planification de la trajectoire des pieds

L'environnement où se déplace la créature doit pouvoir être complexe comme le montre la figure 5.9, avec plusieurs types d'obstacles, ce qui signifie une énorme quantité de triangles pour décrire toutes les géométries du sol et des objets. Sa représentation pour calculer la locomotion est unifiée et simplifiée en utilisant une carte de hauteurs qui peut être construite à partir d'un rendu du dessus par la carte graphique. L'environnement 3D est discrétisé en une grille 2D comme illustrée par la carte d'obstacles à gauche de la figure 5.10 : en gris les zones navigables pour les pieds et en noir les obstacles. La carte d'élévations contient l'élévation de l'obstacle le plus haut dans chaque cellule.

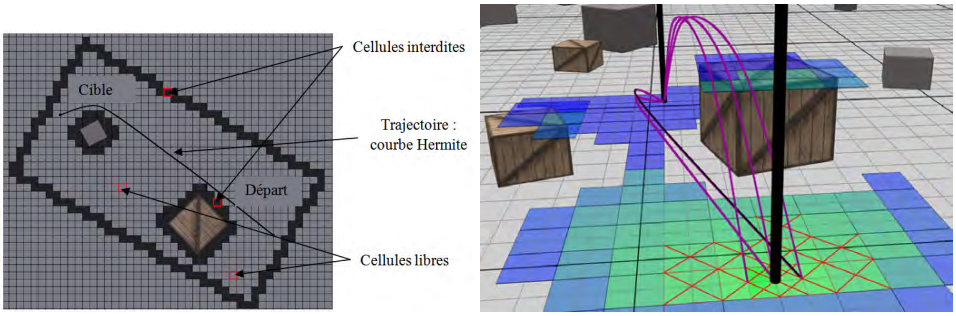


Figure 5.10 : À gauche : la carte d'obstacles vue du dessus. À droite : les cibles potentielles ont une couleur variant du vert (meilleure cible) au bleu (pire cible). Les cellules marquées d'une croix sont les cellules traitées par l'algorithme, en rose les trajectoires possibles et en noir la trajectoire choisie. Source : [Kar+13]

La trajectoire est d'abord calculée en 2D sur le plan horizontal à l'aide des projections de la source et de la cible du pied sur ce plan. Le planificateur discrétise le plan et trouve le plus court chemin avec l'algorithme *WaveFront* [HS99], couplé avec une courbe B-spline. En échantillonnant cette courbe 2D et en utilisant la carte de hauteurs, plusieurs trajectoires

3D sont générées comme le montre la figure 5.10 à droite. Chaque trajectoire 3D peut contourner ou passer au-dessus de chaque obstacle, ce qui génère de multiples trajectoires possibles vers une seule cible. En plus, l'algorithme évalue toutes les cibles (empreintes de pied) autour de l'empreinte cible désignée par le contrôle de la créature. L'espace de recherche est l'ensemble des trajectoires possibles qui vont du point de départ vers toutes les cibles possibles. L'algorithme choisit le meilleur couple (empreinte, trajectoire), en temps réel, avec des critères de longueurs de courbes et de distances à la cible. L'algorithme utilise des scores partiels des trajectoires et des cibles pour limiter le nombre de solutions explorées. Le temps d'exécution de cet algorithme est facilement contrôlable, ce qui permet d'implémenter des techniques de niveau de détail (*Level of Detail - LOD*) pour accélérer les calculs pour les créatures loin ou derrière la caméra afin de garantir le temps réel pour des centaines de créatures.

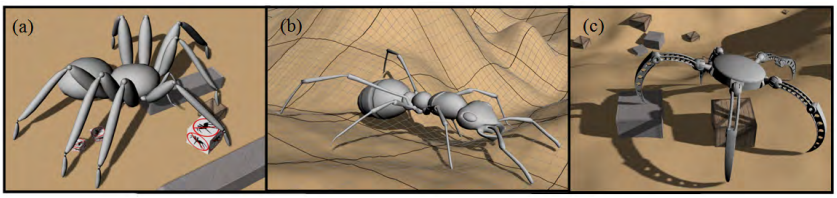


Figure 5.11 : L'approche procédurale permet d'animer facilement différents types de morphologies. Source : [Kar+13]

5.2.4 Colonne vertébrale flexible

La trajectoire du bassin lors de la locomotion chez les humains [Win83] et la plupart des animaux [Muy12] est proche d'une sinusoïde. Pour générer automatiquement ce mouvement sinusoïdal, une approche physique simplement en 1D est proposée. Une particule est utilisée pour représenter le bassin. Son mouvement vertical est régi par la force de gravité qui pousse vers le bas et les forces des pieds qui poussent vers le haut. Chaque pied est traité indépendamment des autres, comme si la particule du bassin se trouvait sur une seule jambe verticale avec un ressort à la place de la jambe, comme un pogostick (bâton sauteur) illustré sur la figure 5.12. Ce pied supporte la masse entière du bassin et pousse vers le haut avec une certaine force pendant la phase au sol.

Les quadrupèdes ont une colonne vertébrale flexible (par ex. les chiens, les lézards), leur conférant une grande agilité, et donc une variété de mouvements étendue. Le bassin de la créature est décomposé en plusieurs nœuds virtuels allant du bassin aux épaules. Chaque pied est relié à l'un de ces nœuds, à l'exception de la tête. Ces nœuds de bassin sont indépendants (hauteur, inclinaison, emplacement des pieds, etc.) et connectés avec notre modèle de colonne vertébrale flexible. Le mouvement de la colonne vertébrale est calculé à chaque pas de temps par quatre étapes, tout d'abord sur le plan horizontal, puis sur le plan sagittal. Sur le plan horizontal, l'algorithme se concentre sur l'orientation 2D et la translation, tandis que sur le plan sagittal, l'algorithme se concentre sur l'élévation avec le système pseudo-physique décrit précédemment et sur l'inclinaison. Dans l'étape finale, les positions 2D et les élévations calculées sont combinées pour obtenir les positions 3D pour chaque nœud du bassin virtuel. Une courbe B-Spline (Hermite) est construite avec ces données pour générer la colonne vertébrale finale comme sur les figures 5.13.

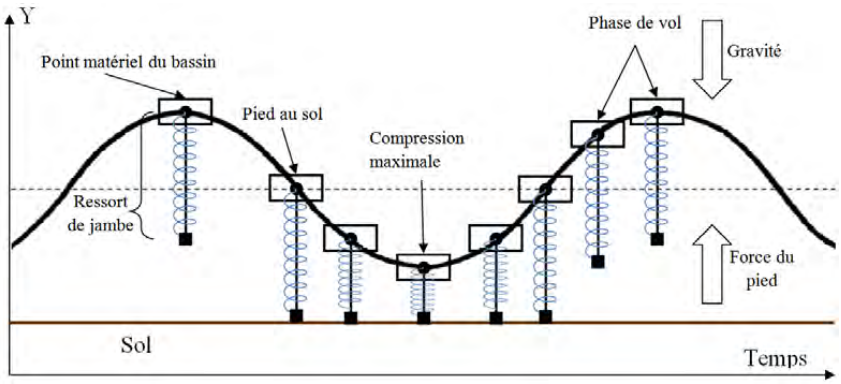


Figure 5.12 : La trajectoire de la particule du bassin est produite par une simulation physique simple en 1D et combine les influences de chaque jambe. Source : [Kar+12]

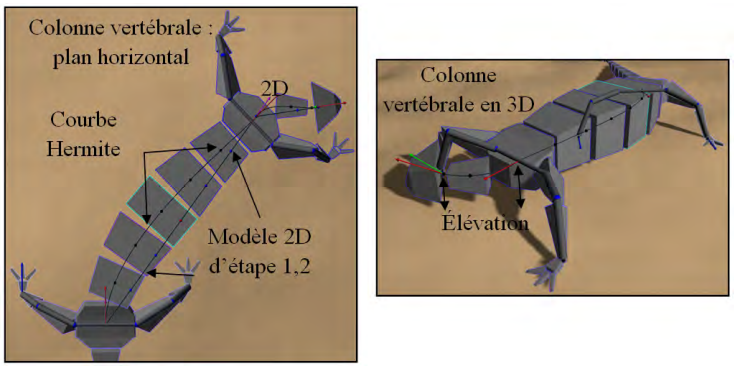


Figure 5.13 : Colonne vertébrale d'un lézard animé procéduralement.

5.2.5 Résultats et conclusion

L'approche présentée ici est capable d'animer, en temps réel, une grande variété de morphologies de créatures à n pattes. Ce système satisfait quatre objectifs principaux. Il est capable d'adapter l'animation générée à un environnement complexe, dynamique et pour différentes morphologies. L'utilisateur a un contrôle total sur la locomotion finale et peut concevoir le style de locomotion désiré à travers les interfaces. Le système génère des animations crédibles et réalistes. Enfin, il est suffisamment efficace pour simuler des dizaines de créatures à n pattes en temps réel.

L'état de l'art s'est éloigné des approches d'animations procédurales au profit des approches basées apprentissage. Pourtant, l'animation procédurale offre des qualités toujours pertinentes : contrôlables, adaptables, explicables, toujours en temps réel, etc. En effet, du fait de la compréhension de chaque partie du processus, il est facile de l'adapter à une nouvelle morphologie, de l'optimiser pour affiner exactement l'algorithme aux besoins ciblés. Cependant, ces approches peuvent souffrir de manque de réalisme quand elles sont appliquées à un humain. Les gestes humains doivent être irréprochables, car l'utilisateur est extrêmement habitué à les observer dans la réalité. Une piste serait de combiner la qualité des animations issues de capture de mouvements aux qualités d'adaptations d'un moteur d'animation procédurale.

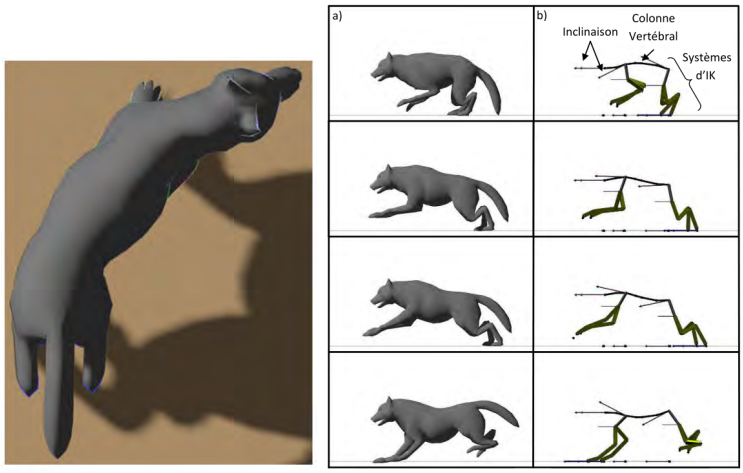


Figure 5.14 : Une animation procédurale de loup avec une colonne vertébrale flexible.

5.3 Édition de poses par apprentissage

Récemment, les approches à base de réseaux de neurones ont proposé des modifications profondes dans les systèmes d'animation en apportant, après apprentissage, des capacités de généralisation à des morphologies différentes [Vil+18 ; Kim+20 ; Lim19 ; Abe+20a], à des environnements variés [HKS17 ; Ber+19 ; LLL18], à des adaptations de gestes et de styles [Hol+17 ; Abe+20b]. Cette section apporte une contribution à cette lignée de travaux "deep" en proposant un système d'édition de poses d'un personnage virtuel basé apprentissage. Le paradigme se propose d'améliorer les systèmes de cinématique inverse (*Inverse Kinematics - IK*) en travaillant sur l'ensemble de la posture du corps. Ces travaux à eux seuls sont une contribution, mais nous argumenterons en conclusion de cette section que ces travaux pourraient être couplés au générateur procédural de trajectoire des pieds de la section précédente pour profiter du meilleur des deux approches.

Dans le pipeline d'animation traditionnel, les animateurs font une large place à l'édition des poses. Ce point est souvent fastidieux, car la cinématique inverse utilisée travaille sur des chaînes cinématiques de manière géométrique sans connaissance des positions possibles et plausibles d'un corps. L'animateur doit donc avoir une compréhension de l'amplitude des mouvements, des possibilités du corps et de ses limites. Cependant, ces informations sont intrinsèquement contenues dans les données de mouvements. Cette section décrit une approche d'édition de poses pilotée par les données, avec pour but d'extraire les connaissances implicites contenues dans des données de mouvements vers un éditeur de pose afin d'en faciliter l'accès aux non initiés. Cette édition de poses s'appuie sur les réseaux neuronaux pour apprendre automatiquement les contraintes à partir des données. La pose est éditée en contrôlant certaines articulations. La section 5.4 décrit une extension pour éditer des descripteurs de style.

5.3.1 Espace latent de poses

L'édition de poses réalisée directement sur les positions ou les rotations des articulations peut être améliorée en concevant un espace d'édition pour les données de pose. Une

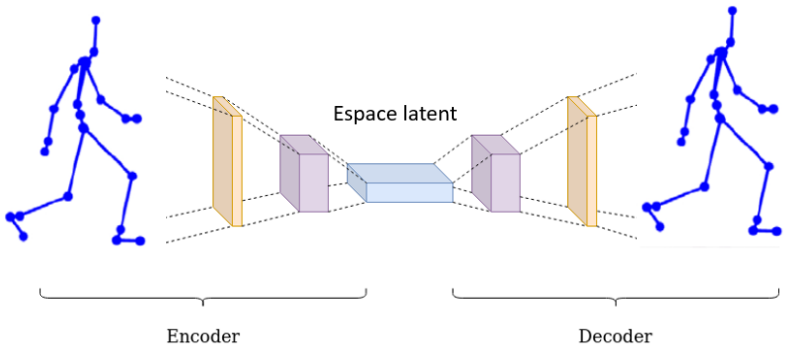


Figure 5.15 : L'auto-encodeur de pose est construit en minimisant l'erreur de reconstruction de la pose d'entrée. Le code produit par l'encodeur correspond à un espace latent qui représente bien l'ensemble des poses plausibles.

approche classique pour construire un espace représentatif des données se base sur une analyse en composante principale comme présentée dans l'approche synthétisant automatiquement une animation neutre dans la section 4.2. Récemment, les réseaux ont gardé cette idée de construction d'un espace alternatif [Bas+21], mais avec plus de capacité à manipuler des données non présentes dans le jeu de données original. D'ailleurs, nos travaux de reconnaissance ont suivi la même évolution avec l'auto-encodeur utilisé dans la section 4.3.2. L'auto-encodeur construit ici travaille sur les poses et non sur l'animation complète comme celui utilisé en reconnaissance d'expressions ou ceux présentés par Holden *et al.* [Hol+15 ; HSK16 ; Hab+17].

Nous avons exploré plusieurs manières de construire un espace latent efficace pour les solveurs d'*IK* : auto-encodeur (*AE*) simple [Hol+15], *Variational-AE* [Hab+17], *AE* et *GAN* combiné (*AEGAN*) [Laz20]. Pour chacun, la génération de points dans l'espace latent permet de garantir que la sortie est toujours une pose plausible, car le décodeur est entraîné à transformer tous les points latents en poses. Finalement, l'auto-encodeur le plus simple est gardé (voir la figure 5.15), car il répond au besoin de l'*IK* et ne nécessite pas des équilibres compliqués à trouver comme dans le cas où l'auto-encodeur et combiné à un *GAN*. Le réseau de l'encodeur est composé de deux couches entièrement connectées avec 200 neurones et des activations *relu* [NH10], suivies d'une couche de sortie sans activation. La taille de la couche de sortie est basée sur le nombre de dimensions dans lesquelles les représentations latentes sont encodées. Nous trouvons empiriquement qu'une dimension de 64 donne un bon équilibre entre la précision de la représentation et la vitesse d'inférence. La dimension du code n'est pas forcément plus petite que les données d'entrées. Simplement, le code représente mieux les données. N'importe quel code donne une pose plausible, ce qui n'est pas le cas en donnant n'importe quelles positions ou rotations aux articulations. Le décodeur est la réplique exacte inversée et utilise le même ensemble de poids. Les poids de l'auto-encodeur sont optimisés en minimisant l'erreur quadratique moyenne entre la pose d'entrée et son équivalent reconstruit.

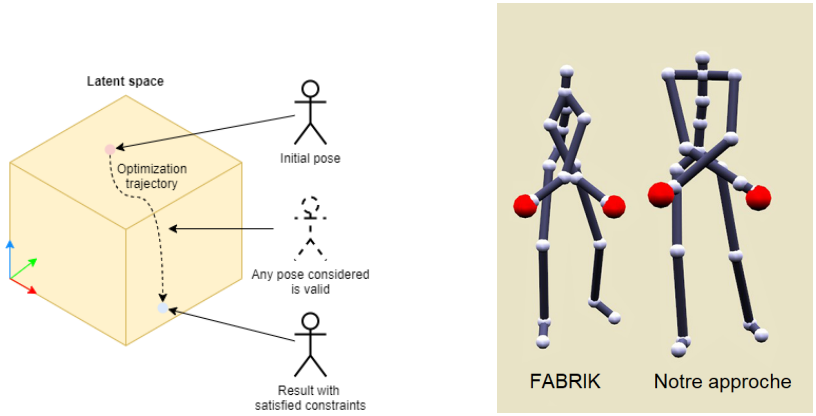


Figure 5.16 : À gauche : l'optimisation dans l'espace latent permet d'éditer une pose pour lui faire atteindre les cibles. À droite : une méthode géométrique d'*IK* comme *FABRIK* [AL11] peut conduire à des poses moins réalistes que notre approche basée sur un apprentissage. Source : [Vic23]

5.3.2 Cinématique inverse dans l'espace latent

L'*IK* du corps entier est une extension du problème original de la cinématique inverse dans laquelle le squelette d'un personnage est considéré dans son intégralité, au lieu d'être un ensemble de chaînes cinématiques. L'objectif est de trouver une position du squelette où une ou plusieurs articulations données sont les plus proches des cibles fournies par l'utilisateur, tout en respectant les contraintes de poses. Les solveurs d'*IK* traditionnels (non basé sur de l'apprentissage) se concentrent sur garantir les longueurs d'os et dans certains cas de garantir les limites d'orientation des articulations. Cependant, les contraintes de pose réelles sont plus subtiles. Le corps doit maintenir un certain équilibre, certaines poses sont inconfortables, voire douloureuses, certaines articulations sont liées, etc. Dans un scénario réel, ces contraintes avancées sont imposées manuellement par les animateurs, avec leur intuition, pour chaque nouveau personnage et chaque nouvelle pose. C'est là que l'espace latent s'avère utile : la fonction de décodage produit des poses compatibles avec l'ensemble des poses des données d'entraînement. La résolution du problème dans l'espace latent garantit que les contraintes de base, comme les contraintes avancées, sont prises en compte. Deux méthodes exploitant l'espace latent sont proposées : la première par optimisation, la seconde avec un autre réseau qui apprend à naviguer dans l'espace latent. Nous ne détaillons que la seconde solution. La solution par optimisation demande plus de temps de calcul et nécessite un espace latent convexe pour pouvoir utiliser un algorithme d'optimisation simple (par exemple descente de gradient) et éviter les minimums locaux. La construction de l'espace latent convexe à base de *VAE* est décrite dans la thèse de Léon Victor [Vic23].

Avec un réseau

Un réseau neuronal est entraîné pour résoudre le problème *IK* dans l'espace latent. La fonction que doit apprendre le réseau est l'optimisation de la somme des distances entre les articulations considérées et leur cible respective. L'idée générale est de modifier le vecteur latent représentant la pose originale pour satisfaire la cible fournie par l'utilisateur. Utiliser

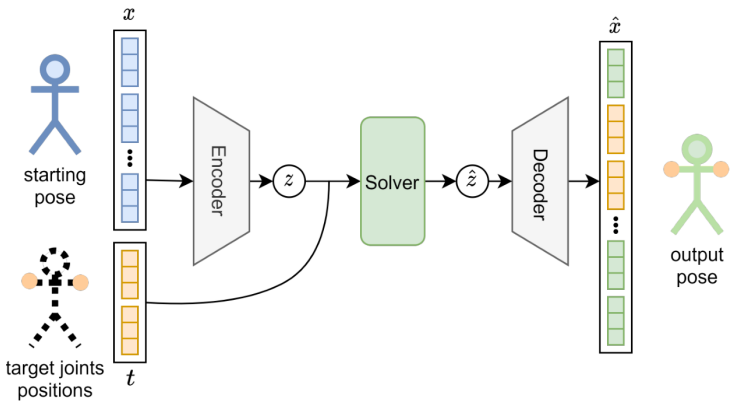


FIGURE 5.17 : Les positions des articulations sources et cibles (jaune) doivent être aussi proches que possible, tandis que les autres articulations (vertes) doivent être aussi proches que possible de la pose de départ (bleu). Source : [VMB21]

un réseau permet d'être plus rapide qu'une optimisation. De plus, comme les réseaux neuronaux sont capables d'apprendre à modéliser des ensembles de données complexes, il est possible de relâcher certaines contraintes sur la construction de l'espace latent. Un auto-encodeur simple comme présenté précédemment est donc suffisant. L'idée est illustrée sur la figure 5.17. Dans un premier temps, le cas d'un unique réseau est considéré. Puis, une méthodologie utilisant plusieurs instances du réseau est présentée pour travailler avec un nombre variable de cibles sans refaire l'entraînement.

Le réseau "solveur de pose" est spécialisé pour résoudre le problème *IK* pour une cible spécifique. Il est entraîné à générer une nouvelle pose à partir d'une pose d'entrée et de l'emplacement des cibles souhaitées. Comme il opère sur l'espace latent construit par l'auto-encodeur, il prend en entrée un code latent et produit un code latent modifié. Le réseau est composé de trois couches entièrement connectées avec 128 neurones, des activations *relu* et d'une couche de sortie dont les dimensions correspondent à celles de l'espace latent. Pendant l'apprentissage, l'ensemble des données est échantillonné aléatoirement pour obtenir une pose d'entrée. Une deuxième pose est échantillonnée à partir du même clip pour l'utiliser comme cible. Nous avons constaté que de rester dans le même clip aide le réseau à mieux apprendre, par rapport à prendre n'importe quel autre clip.

Les poids du réseau sont optimisés pour minimiser la fonction d'erreur dont l'objectif est d'atteindre les cibles des articulations associées tout en conservant une pose réaliste. Il y a donc deux termes dans la fonction d'erreur dont le détail mathématique est disponible dans [VMB21]. Cette fonction d'erreur utilise une erreur quadratique moyenne légèrement modifiée, séparant chaque pose en deux ensembles d'articulations : les articulations associées aux cibles et les autres. Un poids est utilisé pour donner une plus grande importance aux articulations devant toucher les cibles et les articulations n'ayant pas de cibles ne sont utilisées que pour faire pencher le résultat final vers une pose plausible. Dans nos expériences, ce poids est fixé à $0,01$. L'idée est que la pose cible échantillonnée n'est pas une vérité absolue à atteindre à tout prix, mais devrait plutôt être un guide, le but principal étant d'atteindre les cibles.

Cibles multiples

Même si ce solveur est conçu pour générer une pose en tenant compte de plusieurs cibles à la fois, il est possible d'utiliser plusieurs instances séquentiellement pour changer plusieurs fois la pose dans l'espace latent. Dans les cas où l'utilisateur souhaite sélectionner un nombre arbitraire de cibles (pour suggérer une position pour une articulation fixe par exemple), nous pouvons combiner les multiples instances en les exécutant en séquence. Cela permet à la méthode d'être plus modulaire, les utilisateurs pouvant activer ou non un réseau de la séquence et les combiner à leur guise. L'auto-encodeur et les solveurs sont pré-entraînés, les calculs lors de l'utilisation sont donc rapides. Si un utilisateur veut entraîner un solveur pour un jeu d'articulations particulier, seul le solveur est à réentraîner en quelques minutes dans nos expériences.

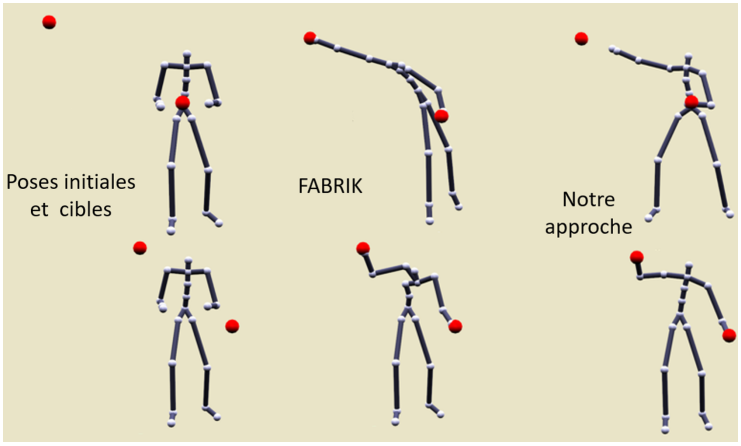


Figure 5.18 : À partir d'une pose et de cibles pour deux articulations, un solveur *IK* comme *FABRIK* (au milieu) génère des poses moins réalistes que notre solveur neuronal (à droite). Source : [VMB21]

5.3.3 Résultats

La figure 5.18 montre un exemple d'édition d'une pose en déplaçant les cibles. La comparaison est réalisée avec *FABRIK*. Le solveur présenté produit des poses satisfaisant les contraintes sans briser les règles implicites du squelette : distances entre articulations constantes, aucune pénétration entre les os et production d'une pose naturelle. Les comparaisons avec *FABRIK* soulignent les limites du travail sur des chaînes cinématiques sans a priori sur le squelette humain. Même si cette méthode s'adresse plutôt aux animateurs débutants, les expérimentés peuvent également la trouver utile. Elle peut par exemple être utilisée comme un outil de prototypage rapide pour initier la pose, puis de passer à une *IK* géométrique dans un deuxième temps.

La figure 5.19 montre un exemple avec une séquence de cinq solveurs : les deux mains, les deux chevilles et la tête. Comparé au résultat de *FABRIK*, la pose est plus plausible : le squelette est plié vers le bas pour atteindre la cible de la tête, mais l'orientation générale de la pose reste intacte. Les membres conservent également une certaine courbure plutôt que de s'étendre complètement de façon non naturelle. Certains objectifs ne sont pas

strictement atteints, car la pose générée par les solveurs précédents est modifiée par les solveurs en fin de séquence, mais la pose obtenue est plausible. L'utilisation d'une séquence de solveurs entraîne des temps d'exécution légèrement plus longs, mais reste plus rapide que *FABRIK*.

Au moment de l'exécution, la complexité du solveur est fixe et, quelle que soit la position des cibles, un seul passage à travers les réseaux suffit à produire un résultat de pose. Les réseaux sont relativement petits et chaque neurone correspond à une multiplication matricielle, ce qui permet un processus de résolution rapide. Par rapport à d'autres méthodes basées sur des données, notre processus nécessite beaucoup de calculs uniquement lors de l'entraînement. Même ainsi, l'entraînement est relativement court : environ une heure pour l'auto-encodeur et 15 minutes pour les solveurs, sur un seul ordinateur. De plus, la quantité de mémoires reste raisonnable en comparaison à des approches basées données qui doivent garder une base de pose en mémoire. Par exemple [WTR11] demande 30 Mo, alors que nos réseaux ne prennent que 826 ko.

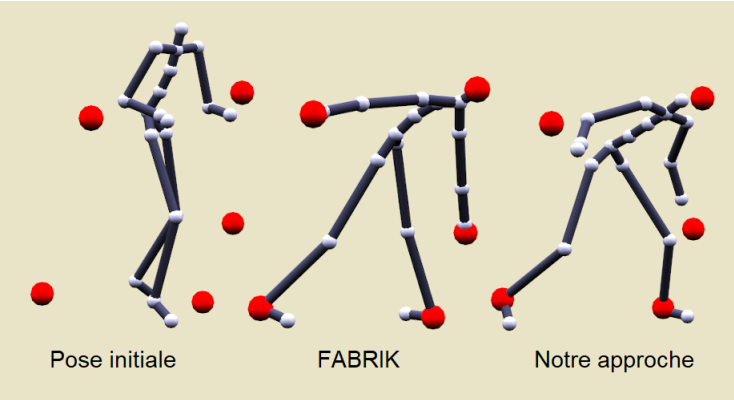


Figure 5.19 : Exemples de résultats de résolution de cibles multiples (ici cinq) avec une séquence de solveurs neuronaux. Les cinq cibles sont indiquées en rouge. Source : [VMB21]

Huang *et al.* [Hua+17] proposent un tableau de comparaison général pour les méthodes d'*IK* du corps entier, classant les approches par rapidité et qualité. L'ajout de notre approche à la figure 5.20 met en évidence la place utile qu'elle occupe en trouvant un bon équilibre entre vitesse et précision. Elle se démarque des approches précédentes basées sur l'apprentissage en étant la première à combiner la vitesse d'édition en temps réel avec des contraintes de squelette entièrement apprises. En comparaison, la méthode *NAT-IK* [Hua+17] utilise des contraintes souples, mais nécessite toujours la définition de contraintes explicites et manuelles. Style-IK [WTR11] ne génère pas les poses en temps réel.

5.3.4 Conclusion et perspective

La méthode proposée exploite les réseaux neuronaux pour fournir un outil interactif et efficace pour éditer le squelette d'un personnage. L'apprentissage à partir d'un grand ensemble de données de poses réelles permet d'éviter de spécifier manuellement les contraintes complexes du squelette humain, et de ne générer que des poses plausibles. L'approche déplace également une grande partie de la charge algorithmique des méthodes traditionnelles vers la phase d'apprentissage, ce qui lui permet de fonctionner à une vitesse

plus que temps réel lors de l'utilisation, en ne demandant qu'une petite quantité de mémoire. Ceci libère des ressources pour d'autres traitements. La suite de ces travaux se trouve dans la section 5.4 où des réseaux seront entraînés pour éditer des descripteurs de styles.

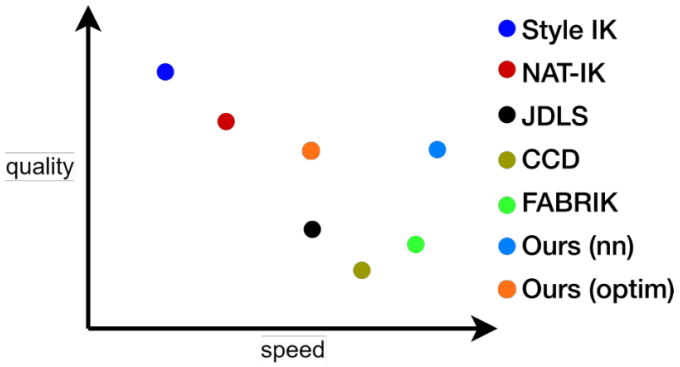


Figure 5.20 : Comparaison générale de diverses méthodes d'IK du corps entier en termes de rapidité et de qualité. Style IK [WTR11], NAT-IK [Hua+17], JDLS [BK05], CCD [WC91a], *FABRIK* [AL11] Source : [Vic23]

Nous aimerions faire un lien avec l'approche procédurale de la section 5.2 qui permet de produire à la volée des animations en temps réel pour n'importe quel environnement. À l'autre bout du spectre de la quantité de données nécessaires, les systèmes d'animations basés sur de l'apprentissage ont d'abord considéré les animations comme une donnée à traiter d'un bloc [HSK16 ; HKS17]. Cependant, pour de nombreux problèmes, les travaux ont eu besoin de proposer des réseaux pour chaque sous-partie du corps. Par exemple, pour la locomotion, Holden *et al.* [Hol+20b] proposent plusieurs réseaux travaillant de concert. Un réseau produit la position et la trajectoire des pieds, puis un ensemble de réseaux produit la posture du corps. Le réseau qui produit la position des pieds garde les défauts de l'apprentissage, à savoir qu'un animateur ne peut pas interagir avec lui pour le contraindre à éviter certaines zones ou en favoriser d'autres sans le refaire un apprentissage. Notre sous-partie d'animation procédurale qui s'occupe de la trajectoire des pieds est très facilement adaptable, car réellement compréhensible. Pour combler les lacunes de l'animation procédurale à générer des poses réalistes et les lacunes de l'apprentissage à être adaptable à la volée sans réapprendre, une idée serait de combiner l'approche procédurale pour la trajectoire des pieds à l'approche apprentissage pour produire les poses réalistes du corps entier.

5.4 Vers des outils contrôlables de production d'animations expressives

L'état de l'art montre que le style est souvent contrôlé à travers des catégories de haut niveau représentant l'état émotionnel de l'avatar (heureux, fier, triste, effrayé, etc.) ou avec des labels de styles (rêveur, âgé, enfantin, épuisé, etc.). Les approches existantes [GEB16 ; Hol+17 ; Jin+19 ; YM16 ; Xia+15b ; Abe+20b] offrent des moyens entièrement automatiques d'application ou de transfert de ces labels à une animation. Ceci peut avoir un intérêt pour

produire une première ébauche d'une animation, mais elles sont ensuite trop restreintes pour que l'animateur puisse exprimer son savoir-faire. Il a besoin de pouvoir contrôler de manière plus fine chaque partie de l'animation. Dans les systèmes d'éditions d'animations à base d'*IK*, il n'y a pas de lien direct avec les styles. L'animateur doit tout réinventer en éditant chaque articulation avec les outils d'éditions classiques. Il manque clairement un niveau intermédiaire d'édition d'animations.

Par exemple, un artiste modifie une séquence d'animations d'un personnage marchant, exprimant une "rage froide". Dans son esprit, le personnage doit faire des pas courts et rapides, les bras tendus vers le sol. Avec les méthodes existantes d'édition de style, il peut appuyer sur un bouton "colère" ou au mieux utiliser un curseur pour déterminer le degré de colère du résultat. Malheureusement, il se peut que le système ait été entraîné avec une autre définition de la colère à l'esprit : le personnage marche maintenant plus vite, les bras pliés, les poings près du torse. L'utilisation de paramètres de pose, bas et haut niveau, permettrait à l'utilisateur de contourner ce problème. En combinant quelques paramètres tels que des angles, des distances entre les extrémités, des paramètres d'effort ou de fluidité, l'animateur pourrait intégrer son savoir-faire. Ces paramètres sont intermédiaires entre le très haut niveau qui correspond juste aux labels et le bas niveau où l'édition se fait par cinématique inverse.

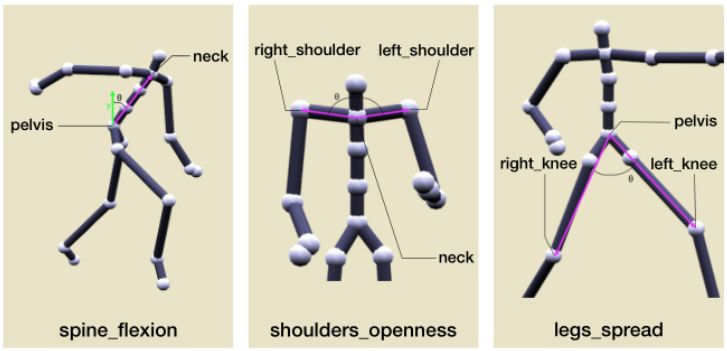


Figure 5.21 : Représentation visuelle des trois caractéristiques de posture éditables par l'utilisateur.

En continuité des travaux décrits précédemment, deux pistes sont actuellement explorées pour arriver à ce but : l'approche par apprentissage et l'approche procédurale. Ce sont des travaux en cours d'élaboration et non publiés, mais il nous semble important d'expliciter ces résultats intermédiaires. Ces solutions ont pour objectif d'offrir à l'utilisateur un panel d'outils d'édition de caractéristiques liées au style. Les caractéristiques identifiées sont soit statiques sur une pose, soit temporelles sur l'animation. Elles sont calculées à partir des données de l'animation et sont fortement liées aux descripteurs identifiés dans la section 4.4 du chapitre sur l'analyse et la reconnaissance des expressions. L'utilisateur doit pouvoir combiner plusieurs de ces paramètres pour les adapter à sa propre vision d'un style. L'approche par apprentissage vise à entraîner des réseaux capables de modifier une animation en contrôlant ces descripteurs expressifs. L'approche procédurale se propose de réaliser cette même tâche avec des outils directs, réellement explicables et contrôlables.

5.4.1 Éditer le style avec des réseaux

La possibilité d'éditer une animation en manipulant des paramètres élaborés, compréhensibles et représentatifs semble une évolution naturelle de la méthode d'édition de poses de la section 5.3. Dans cette partie, l'objectif est de proposer une approche utilisant des réseaux capables de modifier une animation afin qu'elle respecte un ensemble de paramètres de style précis. Pour commencer, trois exemples de métriques applicables à une animation de squelette humain sont proposés : flexion de la colonne vertébrale, ouverture des épaules et écartement des jambes. Ces trois mesures sont illustrées sur la figure 5.21. Il faudrait introduire d'autres paramètres pour prouver que ce processus est un bon outil. Le paramètre de flexion de la colonne vertébrale représente l'angle entre le vecteur "haut" du monde et l'axe de la colonne vertébrale du squelette. La métrique d'ouverture des épaules décrit l'ouverture du torse/épaules du personnage. Des exemples d'utilisations pourraient être de relever les épaules du personnage pour indiquer l'anxiété ou la timidité, et de les ouvrir pour indiquer un comportement plus détendu ou fier. La dernière mesure est l'écartement des jambes qui indique à quel point les genoux du personnage sont proches l'un de l'autre. Parmi les utilisations possibles, il y a la restriction ou l'amplitude des foulées de marche ou la modification de l'apparence d'équilibre du personnage.

L'idée est similaire au solveur de cinématique inverse abordé dans le chapitre précédent : entraîner un réseau neuronal pour générer une nouvelle pose latente à partir d'une pose de départ et d'une valeur cible pour un paramètre de pose spécifique. Si plusieurs paramètres sont édités, une séquence de réseaux neuronaux est appliquée. Une instance d'un réseau est spécialisée pour un seul paramètre de pose, et elle opère dans l'espace latent de pose pour gérer implicitement les contraintes squelettiques.

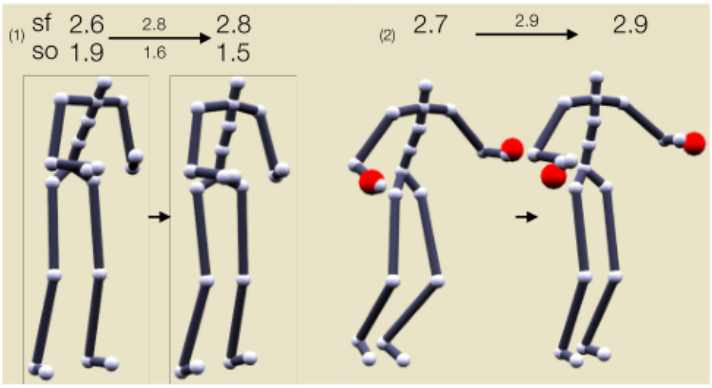


Figure 5.22 : Exemples de résultats du pipeline utilisant plusieurs modules. À gauche, l'ouverture des épaules et la flexion de la colonne vertébrale sont utilisées simultanément. À droite, le solveur d'*IK* du chapitre précédent est utilisé conjointement avec le manipulateur de paramètres de flexion de la colonne vertébrale.

Pour entraîner un tel réseau, deux poses sont échantillonnées dans la même séquence d'animation originale. Le descripteur est calculé sur ces deux poses. L'échantillonnage à partir d'un même clip empêche la pose cible d'être trop différente de la pose de départ. Cela garantit que le réseau apprend à modifier une pose plutôt qu'à en inventer une nouvelle. Le réseau est petit : un perceptron à deux couches de 128 unités cachées et une activation

relu. Le nombre d'unités de la couche d'entrée est égal à la taille de la dimension latente de l'auto-encodeur plus un, pour tenir compte de la valeur du paramètre. La sortie est de même taille que la dimension latente.

Édition temporelle

Le pipeline présenté précédemment n'opère que sur une seule pose à la fois. Cependant, dans un scénario réel, les animateurs souhaitent souvent que leurs changements soient appliqués progressivement aux poses voisines. Pour ceci, la méthode d'édition est étendue à une séquence d'animation entière. L'approche proposée s'inspire du paradigme traditionnel de la courbe d'influence [MK09] afin de propager l'impact d'une modification sur chaque pose voisine avec un poids décroissant. Pour éditer une animation, l'utilisateur sélectionne une pose clé, règle interactivement les valeurs cibles des caractéristiques et la courbe d'influence qu'il souhaite. Cette courbe définit une fenêtre autour de la pose principale sur laquelle les paramètres sont propagés. Chaque pose de la fenêtre est ensuite traitée indépendamment. La fonction d'influence est une fonction chapeau qui culmine à un sur la pose sélectionnée et atteint zéro aux extrémités.

Résultats et perspectives

La figure 5.22 présente des résultats obtenus en éditant plusieurs caractéristiques en séquence. Le premier exemple montre le résultat de la modification simultanée de la flexion de la colonne vertébrale et de l'ouverture des épaules : la colonne vertébrale est redressée et les épaules sont abaissées. Le reste de la pose est également légèrement modifié pour tenir compte de ces changements : les mains sont plus droites et les épaules plus basses. La seconde image montre le résultat de l'utilisation du module de paramètres de flexion de la colonne vertébrale avec le solveur *IK* pour les deux mains présentées précédemment. La méthode permet de trouver un compromis entre chaque contrainte fournie : le dos du squelette est redressé, mais les mains sont abaissées pour compenser la différence, tout en atteignant les cibles.

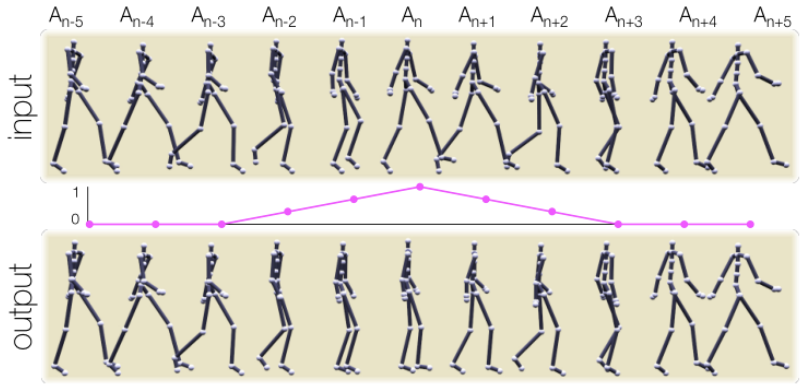


Figure 5.23 : Édition d'un clip d'animation sur une fenêtre. La sortie est produite par le pipeline en fonction de la pose sélectionnée, d'une valeur cible pour la caractéristique d'écartement des jambes et d'une valeur pour la courbe d'influence sur la fenêtre temporelle.

La figure 5.23 montre le résultat d'une animation éditée à l'aide du pipeline d'animation. Sur la pose sélectionnée, la modification de l'angle des jambes est très visible et diminue sur les poses voisines. L'interpolation entre les poses originales et modifiées permet à l'ensemble de la séquence de conserver sa cohérence temporelle, même si le processus travaille pose par pose. Le traitement profite de l'espace latent, car l'interpolation dans l'espace latent est ramenée dans l'espace réel par le décodeur, ce qui permet d'éviter les artefacts tels que l'interpénétration des squelettes.

Il faut considérer ces travaux comme devant être améliorés en ajoutant des descripteurs plus abstraits comme ceux de Laban pour avoir une valeur ajoutée par rapport aux approches d'éditions d'animations existantes [MK09 ; Gua+15]. Nos essais actuels donnent des résultats mitigés. Le style peut apparaître, mais en cassant le réalisme du geste. Probablement, parce que la gestion de l'aspect temporel est ajoutée de manière un peu artificielle à l'espace latent. De plus, la combinaison de plusieurs réseaux donne souvent des modifications contradictoires. Il va falloir intégrer le côté temporel au moment de la construction de l'espace latent. Nous pensons tester par exemple un système de réseaux à deux niveaux comme les deux niveaux *ACP* proposés dans la section 4.2.1 : un espace latent des poses et un espace latent d'une animation complète.

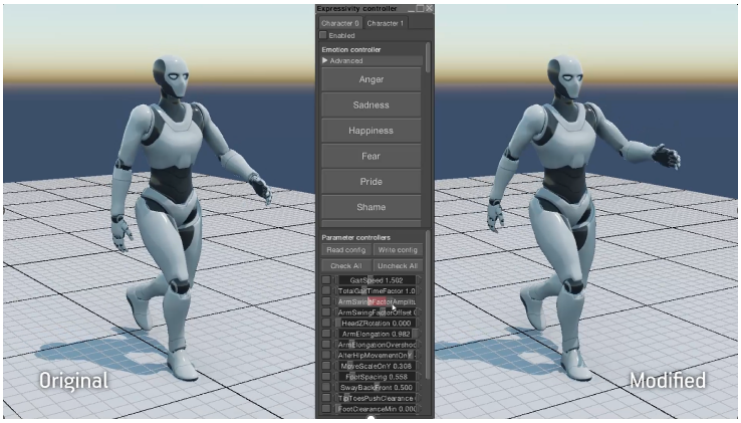


Figure 5.24 : La société *Spirops* avec laquelle nous travaillons dispose d'un moteur d'animation procédurale comportant des dizaines de paramètres pour modifier à la volée une animation de marche.

5.4.2 Éditer le style procéduralement

L'approche d'animation procédurale de la section 5.2 propose une solution intéressante pour la locomotion contrôlable, temps réel et plausible de tous types de créatures à n pattes, sauf pour le bipède humain. Pour la locomotion humaine, le côté plausible est plus difficile à atteindre. Il manque de nombreux paramètres pour reproduire les subtilités de la locomotion humaine qu'un utilisateur a l'habitude d'observer : balancement des bras, colonne vertébrale verticale, oscillation de la tête, respect d'un certain rythme et synchronisations entre articulations, etc. La société *Spirops* a continué de travailler sur le moteur d'animation procédurale et a intégré bon nombre de ces paramètres [Spi22]. L'approche procédurale est plus contrôlable que les techniques de locomotion à base d'apprentissage [HKS17 ; Mas+18]. Même si l'écrasante majorité des publications actuelles

travaillent sur des approches basées sur un apprentissage, il serait dommage de ne pas explorer un peu plus loin les approches procédurales. Le moteur d'animation de *Spirops* permet d'adapter à la volée et interactivement une locomotion humaine en modifiant les paramètres liés à la marche. Même si le réglage de ces paramètres reste complexe, ils peuvent également être déduits d'une animation issue de capture de mouvements pour partir d'une animation réaliste. La figure 5.24 montre un exemple de l'interface où un animateur peut exprimer son savoir-faire et modifier différents paramètres de locomotions. À notre connaissance, il n'existe que deux moteurs d'animations procédurales pour la locomotion utilisé en production, celui d'Ubisoft [Ber16] et celui de *Spirops*.

Les paramètres du moteur d'animation de *Spirops* sont liés à la marche, il n'y a aucun lien entre les styles et ces paramètres. Il y a donc ici une problématique similaire à celle abordée précédemment avec des réseaux dans la section 5.4, mais dans le cadre de l'animation procédurale. Nous basons nos travaux sur les valeurs numériques issues de la méta-analyse de la section 4.4.

Nous avons réalisé des premiers essais en partant d'une publication de Chi *et al.* [Chi+00]. Leur idée est de représenter les angles des articulations de bras par des ellipses paramétrables. Une optimisation sur les paramètres des ellipses permet de faire tendre les descripteurs de styles proposés dans la section 4.4 vers leur valeur chiffrée. Les premiers résultats sont encourageants comme le montre la figure 5.25, mais l'optimisation n'est pas temps réel et les styles obtenus dans les animations sont encore grossiers. Nous travaillons actuellement à améliorer ce point, en examinant tous les détails des descripteurs, notamment leur corrélation.

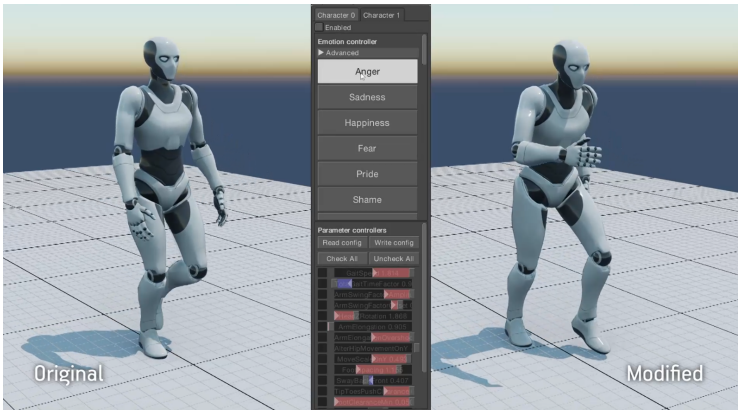


Figure 5.25 : Les paramètres de locomotion procédurale sont optimisés par une descente de gradient pour tendre vers les paramètres de descripteurs de style issus de la méta-analyse de la section 4.4. Ces paramètres sont également éditables à la main pour affiner le résultat. Une animation de marche neutre (à gauche) est automatiquement modifiée pour illustrer la colère à droite : l'amplitude des bras est agrandie, le buste est penché pour produire le côté "décidé", etc.

5.5 Conclusion

Dans ce chapitre, nous avons proposé des approches assez diverses pour l'animation d'humains ou de créatures virtuelles. Plusieurs familles de méthodes sont abordées : animation

de visages basée capture, locomotion procédurale de créatures virtuelles à n pattes, édition de poses par apprentissage, édition de caractéristiques de style par apprentissage et procéduralement. Pour les visages, une technique simple de capture de carte de normales par *Shape from Shading* se basant sur une unique webcam est proposée en association à une approche permettant son transfert sur un visage de personnage virtuel. Pour l'animation temps réel de créatures génériques à n pattes, une technique procédurale de locomotion est proposée en définissant plusieurs modules : un module calculant la position "idéale" des pieds au sol, un module calculant la trajectoire en vol des pieds en évitant les obstacles et un module de colonne vertébrale procédurale. Pour l'édition de poses, nous avons proposé une approche qui se base sur des données et des réseaux de neurones profonds. La méthode repose sur la construction d'un espace de pose latent sur lequel les poses existantes peuvent être projetées puis éditées en respectant les contraintes subtiles qu'un squelette humain doit respecter afin d'être plausible. Un deuxième réseau apprend à éditer la pose dans l'espace latent. Une fois les réseaux entraînés, tout le processus d'édition est temps réel et permet à un animateur d'éditer une pose ou une séquence de poses en manipulant soit les extrémités, soit des descripteurs d'un peu plus haut niveau. À la fin du chapitre, nous ouvrons des perspectives pour proposer des descripteurs plus élaborés, ainsi que l'édition de manière procédurale pour s'affranchir de la phase d'apprentissage des réseaux.

Publications liées

Animation de visages

- Ludovic Dutreve, Alexandre Meyer et Saïda Bouakaz. "Feature points based facial animation retargeting". In : *Proceedings of the 2008 ACM symposium on Virtual reality software and technology*. 2008, p. 197-200
- Ludovic Dutreve, Alexandre Meyer, Veronica Orvalho et Saïda Bouakaz. "Easy rigging of face by automatic registration and transfer of skinning parameters". In : *International Conference on Computer Vision and Graphics*. Springer. 2010, p. 333-341
- Ludovic Dutreve, Alexandre Meyer et Saïda Bouakaz. "Easy acquisition and real-time animation of facial wrinkles". In : *Computer Animation and Virtual Worlds* 22.2-3 (2011), p. 169-176

Animation du corps

- Yann Savoye et Alexandre Meyer. "Multi-Layer Level of Detail For Character Animation". In : *Proceedings of the Fifth Workshop on Virtual Reality Interactions and Physical Simulations, VRIPHYS 2008, Grenoble, France, 2008*. Sous la dir. de François Faure et Matthias Teschner. Eurographics Association, 2008, p. 57-66. url : https://doi.org/10.2312/PE/vriphys/vriphys08/057-066
- Ahmad Abdul Karim, Alexandre Meyer, Thibaut Gaudin, Axel Buendia et Saida Bouakaz. "Generic Spine Model with Simple Physics for Life-Like Quadrupeds and Reptiles." In : *VRIPHYS*. 2012, p. 97-106
- Ahmad Abdul Karim, Thibaut Gaudin, Alexandre Meyer, Axel Buendia et Saida Bouakaz. "Procedural Locomotion of Multilegged Characters in Dynamic Environments". In : *Computer Animation and Virtual Worlds* 24.1 (2013), p. 3-15

- Ahmad Abdul Karim, Thibaut Gaudin, Alexandre Meyer, Axel Buendia et Saïda Bouakaz. “Adding Physical Like Reaction Effects to Skeleton-Based Animations Using Controllable Pendulums”. In : *CASA 2011, Trans. Edutainment* 6 (2011), p. 111-121. URL : https://doi.org/10.1007/978-3-642-22639-7%5C_12
- Léon Victor, Alexandre Meyer et Saïda Bouakaz. “Learning-based pose edition for efficient and interactive design”. In : *Computer Animation and Virtual Worlds* 32.3-4 (2021), e2013

Thèses

- Ludovic Dutreve. “Paramétrisation et Transfert d’Animations Faciales 3D à partir de Séquences Vidéo : vers des Applications en Temps Réel.” Theses. Université Claude Bernard Lyon 1, LIRIS, 2011
- Ahmad Abdul Karim. “Procedural Locomotion of Multi-Legged Characters in Complex Dynamic Environments : Real-Time Applications”. Theses. Université Claude Bernard Lyon 1, LIRIS, 2014
- Léon Victor. “Learning-Based Interactive Character Animation : Expressing Emotions through Motion”. Theses. INSA de Lyon, LIRIS, 2023
- Mehdi-Antoine Mahfoudi. “Génération procédurale d’animations porteuses d’expressions : approche temps réel et interactive”. Theses. Université Claude Bernard Lyon 1, LIRIS, 2023

6 Conclusion et perspectives

Plusieurs communautés proposent des centres d'intérêt qui sont liés à l'étude du geste expressif : la vision par ordinateur, la synthèse d'images, l'IHM, etc. Nos travaux se trouvent au carrefour de ces communautés et se concentrent sur l'analyse, la reconnaissance et la synthèse de mouvements porteurs d'une expression ou d'un style. Nous pensons que ces communautés ont beaucoup en commun même si souvent les contraintes peuvent différer : en vision on privilégie des approches toutes automatiques alors qu'en synthèse et en IHM on garde une large place à l'humain qui doit pouvoir s'exprimer et être compris.

Le chapitre 3 décrit nos contributions en matière de reconnaissance des expressions faciales. Nous avons réalisé une étude perceptive du système visuel humain afin de proposer une nouvelle méthode de reconnaissance d'expressions faciales temps réel et robuste à la dégradation de la résolution des images. Dans une deuxième contribution, nous avons créé une nouvelle base de données pour l'étude des expressions spontanées chez les enfants, catégorie de personne peu représentée dans les données habituellement. Ces travaux sont associés à une technique de reconnaissance mise au point par transfert d'apprentissage.

Le chapitre 4 présente des méthodes pour la reconnaissance des expressions posturales. La reconnaissance est une contribution, mais dans l'idéal nous souhaitons créer des ponts avec la synthèse d'animations. Par exemple, nous proposons plusieurs techniques issues de la synthèse pour générer automatiquement un geste neutre afin d'utiliser leurs différences comme informations discriminantes. Nous proposons également une méthode utilisant un auto-encodeur issu de la synthèse couplé à la matrice de Gram qui obtient des taux de classifications équivalents. Pour finir, nous proposons une formalisation qualitative et quantitative de descripteurs experts efficaces pour la reconnaissance, qui sont aussi pertinents en synthèse d'animations.

Dans le chapitre 5, nous abordons plusieurs techniques d'animation. La première concerne la capture, le transfert et l'animation de visages en temps réel fonctionnant avec une simple webcam. Nous définissons également un système paramétrable d'animation procédurale générique de créatures à n pattes. Des centaines de créatures animées peuvent se déplacer en temps réel dans un environnement dynamique complexe. Puis, en définissant une série de plusieurs petits réseaux de neurones spécialisés, nous introduisons une manière d'éditer une pose humaine restant toujours réaliste et plausible. L'apprentissage profond permet d'intégrer les contraintes de postures humaines dans l'espace latent. Enfin, nous utilisons les descripteurs d'expressions corporelles définis en reconnaissance dans deux méthodes d'éditions d'animations expressives : par apprentissage et procéduralement.

Les données

Nous avons proposé une méthode accessible d'animation de visages en capturant nos propres données à l'aide d'une simple webcam. Pour la reconnaissance d'expressions du visage, nous avons d'abord utilisé les données de visages existantes, puis nous avons

constitué notre propre base de données d'expressions spontanées de visages d'enfants, car cette spécificité est peu présente dans les données disponibles. Notre système d'animation procédurale ciblant des créatures imaginaires ou des insectes (fourmis ou araignées) ne nécessite aucune donnée. L'objectif est de pouvoir contrôler tous les paramètres pour adapter le style de locomotion à chaque créature. Pour l'animation d'humain virtuel comportant différentes expressions, nous explorons deux familles de techniques qui sont diamétralement opposées sur le point des données : à base de réseaux de neurones et procédurale. Même dans le cas de l'apprentissage profond, nous cherchons à rester compétitifs en temps de calcul (moins d'une heure) et en taille de données.

Nous pensons que les données peuvent améliorer les résultats de nombreux problèmes, mais qu'il ne faut pas tout laisser reposer sur leur disponibilité. L'apprentissage profond est un outil puissant, mais il faut le contraindre à aider l'utilisateur en lui laissant une place pour s'exprimer. Un expert peut guider la méthode pour l'améliorer, mais il est aussi important que la méthode lui propose des actions compréhensibles et explicables. Nous pensons que l'apprentissage profond n'est pas incompatible avec ce principe.

Pour aller plus loin...

La reconnaissance des expressions faciales est un domaine qui commence à être assez abouti. Il reste cependant de nombreuses spécificités à explorer, notamment autour des micro-expressions. Les micro-expressions qui sont des expressions qui durent généralement moins d'une seconde et qui révèlent inconsciemment les sentiments d'une personne. Elles sont à peine perceptibles pour un humain. Des travaux existent [Ben+21 ; Aou+21] mais l'hyper-spécificité du problème fait que les experts sont peu nombreux pour aider à la conception des descripteurs et les données sont peu disponibles.

Autour de la reconnaissance des expressions corporelles, nous avons remarqué que le domaine n'effectue jamais de validation croisée sur les sujets : apprendre avec les gestes d'une sous-partie des sujets et tester sur les sujets restants. En comparaison, le domaine de la reconnaissance d'actions fait souvent une validation inter-sujets [WHK20 ; Qin+22] qui donne un meilleur critère sur la capacité de généralisation. Dans un futur proche, nous proposons d'améliorer l'évaluation des techniques de reconnaissance d'expressions par une métrique inter-sujets. Il sera difficile d'obtenir les chiffres de l'état de l'art sur ce point, mais il faudrait que la communauté intègre ce critère.

Le domaine de la reconnaissance pourrait être étendu aux expressions et micro-expressions du corps et des mains. Une personne réalise beaucoup d'actions avec ses mains [Bou+22] et exprime donc des émotions. De plus, les techniques de capture de la main incluant les doigts sont maintenant utilisables avec une seule caméra [Lug+19]. Le domaine de la reconnaissance d'actions des mains dispose déjà de techniques de reconnaissance [Bou+22]. Pour les expressions, il serait intéressant de combiner les mouvements des doigts, des mains, du corps et du visage. De plus, autant pour le corps que pour le visage, il reste de nombreuses expressions plus subtiles à étudier, comme en témoigne la diversité d'adjectifs existants [VBP20 ; Aud23]. Il serait intéressant également de transposer nos travaux en considérant un espace des émotions continu qui est plus pratique en synthèse d'animations. Il reste de nombreux travaux à effectuer sur l'identification de biais présent dans les bases

de données [LD20]. Il faudrait être capable de les identifier et de proposer des approches pour les corriger, sans forcément recourir à l'obtention de nouvelles données.

L'approche proposée au chapitre 4 qui sépare l'action de l'expression peut mener à de nouvelles applications en synthèse d'animations, car elle est compréhensible. Il faudrait notamment la transférer dans le monde des réseaux. En effet, les réseaux ont une capacité à généraliser qui est probablement plus forte que l'*ACP* que nous avons utilisée. Pour cela, utiliser les réseaux d'attentions semble une piste intéressante, car ils trouveront des corrélations entre des articulations séparées dans le squelette. Les réseaux avec une topologie de graphe [Lia+22] ont montré leur efficacité dans de nombreux cas, mais les réseaux d'attentions devraient être plus efficaces lorsque les deux mains se touchent ou réalisent une action coordonnée. Ces réseaux sont à investiguer aussi bien en reconnaissance, qu'en synthèse. Les réseaux de diffusion proposent également de jolis résultats en génération d'images avec par exemple *Dall-E* [Ram+21b]. Des publications commencent à émerger en combinant le texte et les animations [Hon+22]. Cette voie semble intéressante pour aider un novice à créer une animation. Mais, il reste de nombreuses pistes à investiguer pour y inclure le visage, le corps, les mains, les doigts, ainsi que tout le vocabulaire des expressions et des styles. Pour les animateurs plus expérimentés, de nombreux outils restent à inventer en combinant toutes les connaissances expertes acquises aux connaissances qui peuvent émerger des données.

Pour conclure, de nombreuses techniques compréhensibles et contrôlables restent à inventer afin d'améliorer la reconnaissance automatique d'expressions et la création d'animations "stylées".

7 Publications

7.1 Revues internationales à comité de lecture

- Mehdi-Antoine Mahfoudi, Alexandre Meyer, Thibaut Gaudin, Axel Buendia et Saida Bouakaz. "Emotion Expression in Human Body Posture and Movement : A Survey on Intelligible Motion Factors, Quantification and Validation". In : *IEEE Transactions on Affective Computing* (2022), p. 1-24

- Léon Victor, Alexandre Meyer et Saïda Bouakaz. "Learning-based pose edition for efficient and interactive design". In : *Computer Animation and Virtual Worlds* 32.3-4 (2021), e2013

- Arthur Crenn, Alexandre Meyer, Hubert Konik, Rizwan Ahmed Khan et Saida Bouakaz. "Generic body expression recognition based on synthesis of realistic neutral motion". In : *IEEE Access* 8 (2020), p. 207758-207767

- Rizwan Ahmed Khan, Arthur Crenn, Alexandre Meyer et Saida Bouakaz. "A novel database of children's spontaneous facial expressions (LIRIS-CSE)". In : *Image and Vision Computing* 83-84 (2019), p. 61-69

- Rizwan Ahmed Khan, Alexandre Meyer, Hubert Konik et Saida Bouakaz. "Saliency-based framework for facial expression recognition". In : *Frontiers of Computer Science* 13.1 (2019), p. 183-198

- Ahmad Abdul Karim, Thibaut Gaudin, Alexandre Meyer, Axel Buendia et Saida Bouakaz. "Procedural Locomotion of Multilegged Characters in Dynamic Environments". In : *Computer Animation and Virtual Worlds* 24.1 (2013), p. 3-15

- Rizwan Ahmed Khan, Alexandre Meyer, Hubert Konik et Saida Bouakaz. "Framework for reliable, real-time facial expression recognition for low resolution images". In : *Pattern Recognition Letters* 34.10 (2013), p. 1159-1168

- Ahmad Abdul Karim, Thibaut Gaudin, Alexandre Meyer, Axel Buendia et Saïda Bouakaz. "Adding Physical Like Reaction Effects to Skeleton-Based Animations Using Controllable Pendulums". In : *CASA 2011, Trans. Edutainment* 6 (2011), p. 111-121. url : https://doi.org/10.1007/978-3-642-22639-7%5C_12

- Ludovic Dutreve, Alexandre Meyer et Saïda Bouakaz. "Easy acquisition and real-time animation of facial wrinkles". In : *Computer Animation and Virtual Worlds* 22.2-3 (2011), p. 169-176

7.2 Actes de conférences internationales à comité de lecture

- Arthur Crenn, Alexandre Meyer, Rizwan Ahmed Khan, Hubert Konik et Saïda Bouakaz. "Toward an Efficient Body Expression Recognition Based on the Synthesis of a Neutral Movement". In : ICMI 2017 Proceedings of the 19th ACM International Conference on Multimodal Interaction (2017)
- Arthur Crenn, Rizwan Ahmed Khan, Alexandre Meyer et Saida Bouakaz. "Body expression recognition from animated 3D skeleton". In : (2016), p. 1-7
- Rizwan Ahmed Khan, Alexandre Meyer et Saida Bouakaz. "Automatic affect analysis : from children to adults". In : *International Symposium on Visual Computing*. Springer. 2015, p. 304-313
- Rizwan Ahmed Khan, Alexandre Meyer, Hubert Konik et Saïda Bouakaz. "Pain detection through shape and appearance features". In : *Proceedings of the 2013 IEEE International Conference on Multimedia and Expo, ICME 2013, San Jose, CA, USA, July 15-19, 2013*. IEEE Computer Society, 2013, p. 1-6
- Ahmad Abdul Karim, Alexandre Meyer, Thibaut Gaudin, Axel Buendia et Saida Bouakaz. "Generic Spine Model with Simple Physics for Life-Like Quadrupeds and Reptiles." In : *VRIPHYS*. 2012, p. 97-106
- Rizwan Ahmed Khan, Alexandre Meyer, Hubert Konik et Saida Bouakaz. "Exploring human visual system : study to aid the development of automatic facial expression recognition framework". In : *2012 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*. IEEE. 2012, p. 49-54
- Rizwan Ahmed Khan, Alexandre Meyer, Hubert Konik et Saida Bouakaz. "Human Vision Inspired Framework for Facial Expressions Recognition". In : *2012 19th IEEE International Conference on Image Processing*. Sept. 2012, p. 2593-2596
- Ludovic Dutreve, Alexandre Meyer, Veronica Orvalho et Saïda Bouakaz. "Easy rigging of face by automatic registration and transfer of skinning parameters". In : *International Conference on Computer Vision and Graphics*. Springer. 2010, p. 333-341
- Ludovic Dutreve, Alexandre Meyer et Saïda Bouakaz. "Feature points based facial animation retargeting". In : *Proceedings of the 2008 ACM symposium on Virtual reality software and technology*. 2008, p. 197-200
- Yann Savoye et Alexandre Meyer. "Multi-Layer Level of Detail For Character Animation". In : *Proceedings of the Fifth Workshop on Virtual Reality Interactions and Physical Simulations, VRIPHYS 2008, Grenoble, France, 2008*. Sous la dir. de François Faure et Matthias Teschner. Eurographics Association, 2008, p. 57-66. URL : https://doi.org/10.2312/PE/vriphys/vriphys08/057-066
- Alexandre Meyer, Hector M Briceno et Saida Bouakaz. "User-guided shape from shading to reconstruct fine details from a single photograph". In : *Asian Conference on Computer Vision*. Springer. 2007, p. 738-747

7.3 Communications à des congrès, symposiums nationaux

- Léon Victor et Alexandre Meyer. "Character Pose Design in Latent Space For Animation Edition". In : *Journées Françaises de l'Informatique Graphique 2020*. Nancy, France, nov. 2020. URL : https://hal.science/hal-03338910
- Arthur Crenn, Hubert Konik, Alexandre Meyer et Saïda Bouakaz. "Reconnaissance d'expressions corporelles à l'aide d'un mouvement neutre synthétisé". In : *CORESA 2017 (COmpression et REpresentation des Signaux Audiovisuels)*. Laboratoire GREYC (UMR CNRS 6072). Caen, France, nov. 2017. URL : https://hal.science/hal-01680819

7.4 Rapports

- Léon Victor, Alexandre Meyer et Saïda Bouakaz. "Pose Metrics : a New Paradigm for Character Motion Edition". In : *CoRR* abs/2301.06514 (2023). URL : https://doi.org/10.48550/arXiv.2301.06514

7.5 Base de données

- R. A. Khan, A. Crenn, A. Meyer et S. Bouakaz. *LIRIS-CSE Children Spontaneous Facial Expression Video Database*. 2020. URL : https://childrenfacialexpression.projet.liris.cnrs.fr (305 chercheurs enregistrés au 20 janvier 2023).

Table des figures

Liste des tableaux

Bibliographie

[AL14] Paul A. Wilson et Barbara Lewandowska-Tomaszczyk. “Affective Robotics : Modelling and Testing Cultural Prototypes”. In : *Cognitive Computation* 6.4 (déc. 2014), p. 814-840 (cf. p. 79).

[Abd14] Ahmad Abdul Karim. “Procedural Locomotion of Multi-Legged Characters in Complex Dynamic Environments : Real-Time Applications”. Theses. Université Claude Bernard Lyon 1, LIRIS, 2014 (cf. p. 12, 13, 110).

[Abe+20a] Kfir Aberman, Peizhuo Li et al. “Skeleton-Aware Networks for Deep Motion Retargeting”. In : *ACM Trans. Graph.* 39.4 (juill. 2020), 62 :62 :1-62 :62 :14 (cf. p. 97).

[Abe+20b] Kfir Aberman, Yijia Weng, Dani Lischinski, Daniel Cohen-Or et Baoquan Chen. “Unpaired Motion Style Transfer from Video to Animation”. In : *ACM Transactions on Graphics* 39.4 (juill. 2020), 64 :64 :1-64 :64 :12 (cf. p. 97, 103).

[Abe+20c] Kfir Aberman, Yijia Weng, Dani Lischinski, Daniel Cohen-Or et Baoquan Chen. “Unpaired motion style transfer from video to animation”. In : *ACM Transactions on Graphics (TOG)* 39.4 (2020), p. 64-1 (cf. p. 48).

[Abi+22] Vicenzo Abichequer Sangalli, Ludovic Hoyet, Marc Christie et Julien Pettré. “A new framework for the evaluation of locomotive motion datasets through motion matching techniques”. In : *Proceedings of the 15th ACM SIGGRAPH Conference on Motion, Interaction and Games*. 2022, p. 1-10 (cf. p. 45).

[AB18] Amina Adadi et Mohammed Berrada. “Peeking Inside the Black-Box : A Survey on Explainable Artificial Intelligence (XAI)”. In : *IEEE Access* 6 (2018), p. 52138-52160 (cf. p. 81).

[APD10] N. Aifanti, C. Papachristou et A. Delopoulos. “The MUG facial expression database”. In : *11th International Workshop on Image Analysis for Multimedia Interactive Services WIAMIS 10*. 2010, p. 1-4 (cf. p. 29).

[AMD19] Insaf Ajili, Malik Mallem et Jean-Yves Didier. “Human motions and emotions recognition inspired by LMA qualities”. In : *The Visual Computer* 35.10 (2019), p. 1411-1426 (cf. p. 37, 80).

[El-+16] Moe El-Ali, Le Tong et al. “Zootopia crowd pipeline”. In : *ACM SIGGRAPH 2016 Talks*. 2016, p. 1-2 (cf. p. 40).

[Alo+18] Md Zahangir Alom, Tarek M Taha et al. “The history began from alexnet : A comprehensive survey on deep learning approaches”. In : *arXiv preprint arXiv :1803.01164* (2018) (cf. p. 59).

[Alv+22] Eduardo Alvarado, Chloé Paliard, Damien Rohmer et Marie-Paule Cani. “Real-Time Locomotion on Soft Grounds With Dynamic Footprints”. In : *Frontiers in Virtual Reality* 3 (2022) (cf. p. 41).

[ABC96] Kenji Amaya, Armin Bruderlin et Tom Calvert. “Emotion from motion”. In : *Graphics interface*. T. 96. 00306. Toronto, Canada, 1996, p. 222-229. url : http://www.graphicsinterface.org/wp-content/uploads/gi1996-26.pdf (visité le 19 nov. 2015) (cf. p. 47).

[Aou+21] Mouath Aouayeb, Wassim Hamidouche, Catherine Soladie, Kidiyo Kpalma et Renaud Seguier. “Micro-expression recognition from local facial regions”. In : *Signal Processing : Image Communication* 99 (2021), p. 116457 (cf. p. 30, 112).

[AF02] Okan Arikan et David A Forsyth. “Interactive motion generation from examples”. In : *ACM Transactions on Graphics (TOG)* 21.3 (2002), p. 483-490 (cf. p. 44).

[ACL16] Andreas Aristidou, Yiorgos Chrysanthou et Joan Lasenby. “Extending FABRIK with model constraints”. In : *Computer Animation and Virtual Worlds* 27.1 (2016), p. 35-57 (cf. p. 43).

[AL11] Andreas Aristidou et Joan Lasenby. “FABRIK : A fast, iterative solver for the Inverse Kinematics problem”. en. In : *Graphical Models* 73.5 (2011). 00104, p. 243-260. url : http://linkinghub.elsevier.com/retrieve/pii/S1524070311000178 (visité le 21 mars 2017) (cf. p. 42, 99, 103).

[Ari+18] Andreas Aristidou, Joan Lasenby, Yiorgos L. Chrysanthou et A. Shamir. “Inverse Kinematics Techniques in Computer Graphics : A Survey : Inverse Kinematics Techniques in Computer Graphics”. In : *Computer Graphics Forum* 37.6 (sept. 2018), p. 35-58 (cf. p. 42).

[Ari+15] Andreas Aristidou, Efstathios Stavrakis, Panayiotis Charalambous, Yiorgos Chrysanthou et Stephania Loizidou Himona. “Folk dance evaluation using laban movement analysis”. In : *Journal on Computing and Cultural Heritage (JOCCH)* 8.4 (2015), p. 1-19 (cf. p. 36).

[AWH92] Joel Aronoff, Barbara A. Woike et Lester M. Hyman. “Which are the stimuli in facial displays of anger and happiness ? : Configurational bases of emotion recognition.” In : *Journal of Personality and Social Psychology* 62.6 (1992), p. 1050-1066 (cf. p. 33).

[AWS19] Ilham Arun Faisal, Tito Waluyo Purboyo et Anton Siswo Raharjo Ansori. “A Review of Accelerometer Sensor and Gyroscope Sensor in IMU Sensors on Motion Capture”. In : *Journal of Engineering and Applied Sciences* 15.3 (nov. 2019), p. 826-829 (cf. p. 30).

[ATD07] A. P. Atkinson, M. L. Tunstall et W. H. Dittrich. “Evidence for distinct contributions of form and motion information to the recognition of emotions from body gestures”. In : *Cognition* 104.1 (2007), p. 59-72 (cf. p. 31).

[Aud23] Daniel Auduc. *Les adjectifs qualificatifs descriptifs*. 2023. url : https://www.danielauduc.fr/e2c/complangue/fle/mots-pour-ecrire/adjectifs%20descriptifs.pdf (cf. p. 112).

[BMS12] T. Banziger, M. Mortillaro et K. R. Scherer. “Introducing the Geneva Multimodal expression corpus for experimental research on emotion perception”. In : *Emotion* 12.5 (2012), p. 1161-1179 (cf. p. 33).

[Bar+13] Avi Barliya, Lars Omlor, Martin A Giese, Alain Berthoz et Tamar Flash. "Expression of emotion in the kinematics of locomotion". In : *Experimental brain research* 225.2 (2013), p. 159-176 (cf. p. 36).

[Bar+12] Mathieu Barnachon, Saïda Bouakaz, Boubakeur Boufama et Erwan Guillou. "Human actions recognition from streamed motion capture". In : *Proceedings of the 21st international conference on pattern recognition (ICPR2012)*. IEEE. 2012, p. 3807-3810 (cf. p. 35).

[Bar+14] Mathieu Barnachon, Saïda Bouakaz, Boubakeur Boufama et Erwan Guillou. "Ongoing human action recognition with motion capture". In : *Pattern Recognition* 47.1 (2014), p. 238-247 (cf. p. 30).

[Bar+06] M.S. Bartlett, G. Littlewort et al. "Fully Automatic Facial Action Recognition in Spontaneous Behavior". In : *7th International Conference on Automatic Face and Gesture Recognition (FGR06)*. Avr. 2006, p. 223-230 (cf. p. 57).

[BV08] Shishir Bashyal et Ganesh K. Venayagamoorthy. "Recognition of facial expressions using Gabor wavelets and learning vector quantization". In : *Engineering Applications of Artificial Intelligence* 21.7 (2008), p. 1056-1064. url : http://www.sciencedirect.com/science/article/pii/S0952197607001492 (cf. p. 22, 23).

[Bas+21] Jean Basset, Adnane Boukhayma, Stefanie Wuhrer, Franck Multon et Edmond Boyer. "Neural Human Deformation Transfer". In : *2021 International Conference on 3D Vision (3DV)*. IEEE. 2021, p. 545-554 (cf. p. 98).

[Bas79] John N Bassili. "Emotion recognition : The role of facial movement and the relative importance of upper and lower areas of the face." In : *Journal of personality and social psychology* 37.11 (1979), p. 2049 (cf. p. 57).

[Ben+21] Xianye Ben, Yi Ren et al. "Video-based facial micro-expression analysis : A survey of datasets, features and algorithms". In : *IEEE transactions on pattern analysis and machine intelligence* (2021) (cf. p. 30, 112).

[BNK18] Gibran Benitez-Garcia, Tomoaki Nakamura et Masahide Kaneko. "Multicultural facial expression recognition based on differences of western-caucasian and east-asian facial expressions of emotions". In : *IEICE TRANSACTIONS on Information and Systems* 101.5 (2018), p. 1317-1324 (cf. p. 24).

[Ber16] Alexander Bereznyak. "IK Rig : Procedural Pose Animation (Ubisoft)". In : *Game Developers Conference. http ://schedule.gdconf.com/session/ik-rig-procedural-pose-animation*. 2016 (cf. p. 108).

[Ber+19] Kevin Bergamin, Simon Clavet, Daniel Holden et James Richard Forbes. "DReCon : data-driven responsive control of physics-based characters". In : *ACM Transactions On Graphics (TOG)* 38.6 (2019), p. 1-11 (cf. p. 97).

[BR07] Daniel Bernhardt et Peter Robinson. "Detecting affect from non-stylised body motions". In : *Affective Computing and Intelligent Interaction*. 00129. Springer, 2007, p. 59-70. url : http://link.springer.com/chapter/10.1007/978-3-540-74889-2_6 (visité le 30 mars 2016) (cf. p. 36, 66).

[BP93] Diane S Berry et James W Pennebaker. “Nonverbal and verbal emotional expression and health”. In : *Psychotherapy and psychosomatics* 59.1 (1993), p. 11-19 (cf. p. 16).

[BP04] Elisabetta Bevacqua et Catherine Pelachaud. “Expressive Audio-Visual Speech”. In : *Computer Animation and Virtual Worlds* 15.3-4 (2004), p. 297-304 (cf. p. 41).

[Bey+21] Cigdem Beyan, Sukumar Karumuri, Gualtiero Volpe, Antonio Camurri et Radoslaw Niewiadomski. “Modeling Multiple Temporal Scales of Full-body Movements for Emotion Classification”. In : *IEEE Transactions on Affective Computing* (2021), p. 1-1 (cf. p. 38, 80).

[Bha+21] Uttaran Bhattacharya, Nicholas Rewkowski et al. “Text2gestures : A transformer-based network for generating emotive body gestures for virtual agents”. In : *2021 IEEE Virtual Reality and 3D User Interfaces (VR)*. IEEE. 2021, p. 1-10 (cf. p. 48).

[Bha+20] Uttaran Bhattacharya, Christian Roncal et al. “Take an emotion walk : Perceiving emotions from gaits using hierarchical attention pooling and affective mapping”. In : *European Conference on Computer Vision*. Springer. 2020, p. 145-163 (cf. p. 36, 38, 48).

[Bha19] Zeeshan Bhatti. “Oscillator driven central pattern generator (CPG) system for procedural animation of quadruped locomotion”. In : *Multimedia Tools and Applications* 78.21 (2019), p. 30485-30502 (cf. p. 40).

[BK03] Nadia Bianchi-berthouze et Andrea Kleinsmith. “A categorical approach to affective gesture recognition”. In : *Connection Science* 15.4 (2003), p. 259-269. eprint : https://doi.org/10.1080/09540090310001658793. url : https://doi.org/10.1080/09540090310001658793 (cf. p. 39).

[Bla49] Preston Blair. *Advanced animation*. Foster, 1949 (cf. p. 43).

[BB18] Adnane Boukhayma et Edmond Boyer. “Surface motion capture animation synthesis”. In : *IEEE transactions on visualization and computer graphics* 25.6 (2018), p. 2270-2283 (cf. p. 91).

[Bou+22] Yasser Boutaleb, Catherine Soladié et al. “Metric Learning-Based Unsupervised Domain Adaptation for 3D Skeleton Hand Activities Categorization”. In : *Image Analysis and Processing–ICIAP 2022 : 21st International Conference, Lecce, Italy, May 23–27, 2022, Proceedings, Part III*. Springer. 2022, p. 62-74 (cf. p. 112).

[Bou+21] Yasser Mohamed Boutaleb, Catherine Soladie et al. “Efficient multi-stream temporal learning and post-fusion strategy for 3D skeleton-based hand activity recognition”. In : *International Conference on Computer Vision Theory and Applications (VISAPP)*. T. 4. Scitepress. 2021, p. 293-302 (cf. p. 30).

[BH00] Matthew Brand et Aaron Hertzmann. “Style machines”. In : *Proceedings of the 27th annual conference on Computer graphics and interactive techniques*. 00740. ACM Press/Addison-Wesley Publishing Co., 2000, p. 183-192. url : http://dl.acm.org/citation.cfm?id=344865 (visité le 26 oct. 2015) (cf. p. 46).

[Bro+97] Sheila Brownlow, Amy R. Dixon, Carrie A. Egbert et Rebecca D. Radcliffe. "Perception of Movement and Dancer Characteristics from Point-Light Displays of Dance". In : *The Psychological Record* 47.3 (juill. 1997), p. 411-422 (cf. p. 79).

[BW95] Armin Bruderlin et Lance Williams. "Motion signal processing". In : *Proceedings of the 22nd annual conference on Computer graphics and interactive techniques*. 00955. ACM, 1995, p. 97-104. url : http://dl.acm.org/citation.cfm?id=218421 (visité le 21 oct. 2015) (cf. p. 47).

[Buh00] Martin D Buhmann. "Radial basis functions". In : *Acta numerica* 9 (2000), p. 1-38 (cf. p. 89).

[Bui+14] Stéphanie Buisine, Matthieu Courgeon et al. "The Role of Body Postures in the Recognition of Emotions in Contextually Rich Scenarios". In : *Int. J. Hum. Comput. Interact.* 30.1 (2014), p. 52-62. url : https://doi.org/10.1080/10447318.2013.802200 (cf. p. 16, 30).

[Bus04] Samuel R Buss. "Introduction to inverse kinematics with jacobian transpose, pseudoinverse and damped least squares methods". In : *IEEE Journal of Robotics and Automation* 17.1-19 (2004), p. 16 (cf. p. 42).

[BK05] Samuel R. Buss et Jin-Su Kim. "Selectively Damped Least Squares for Inverse Kinematics". In : *Journal of Graphics Tools* 10.3 (jan. 2005), p. 37-49 (cf. p. 103).

[Bus+07] Carlos Busso, Zhigang Deng, Michael Grimm, Ulrich Neumann et Shrikanth Narayanan. "Rigid Head Motion in Expressive Speech Animation : Analysis and Synthesis". In : *IEEE Transactions on Audio, Speech, and Language Processing* 15.3 (mars 2007), p. 1075-1086 (cf. p. 77, 79).

[CB94] John T Cacioppo et Gary G Berntson. "Relationship between attitudes and evaluative space : A critical review, with emphasis on the separability of positive and negative substrates." In : *Psychological bulletin* 115.3 (1994), p. 401 (cf. p. 20).

[Cai+21] Jie Cai, Zibo Meng et al. "Identity-free facial expression recognition using conditional generative adversarial network". In : *2021 IEEE International Conference on Image Processing (ICIP)*. IEEE. 2021, p. 1344-1348 (cf. p. 28).

[Cam+11] A. Camurri, N. Dael et al. In : *IEEE Transactions on Affective Computing* 2.02 (2011), p. 106-118 (cf. p. 33).

[Cam+03] Antonio Camurri, Barbara Mazzarino, Matteo Ricchetti, Renee Timmers et Gualtiero Volpe. "Multimodal analysis of expressive gesture in music and dance performances". In : *International gesture workshop*. Springer. 2003, p. 20-39 (cf. p. 36).

[CTV02] Antonio Camurri, Riccardo Trocca et Gualtiero Volpe. "Interactive Systems Design : A KANSEI-based Approach". In : *Proceedings of the 2002 Conference on New Interfaces for Musical Expression*. NIME '02. Dublin, Ireland : National University of Singapore, 2002, p. 1-8. url : http://dl.acm.org/citation.cfm?id=1085171.1085208 (cf. p. 36).

[Can+22] Felipe Zago Canal, Tobias Rossi Müller et al. “A survey on facial emotion recognition techniques : A state-of-the-art literature review”. In : *Information Sciences* 582 (2022), p. 593-617. url : https://www.sciencedirect.com/science/article/pii/S0020025521010136 (cf. p. 22, 27, 28).

[Cao+21] Chen Cao, Vasu Agrawal et al. “Real-time 3D neural facial animation from binocular video”. In : *ACM Transactions on Graphics (TOG)* 40.4 (2021), p. 1-17 (cf. p. 91).

[Cao+19] Z. Cao, G. Hidalgo Martinez, T. Simon, S. Wei et Y. A. Sheikh. “OpenPose : Realtime Multi-Person 2D Pose Estimation using Part Affinity Fields”. In : *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2019) (cf. p. 30, 35).

[CK73] William R Charlesworth et Mary Anne Kreutzer. “Facial expressions of infants and children”. In : *Darwin and facial expression : A century of research in review* (1973), p. 91-168 (cf. p. 57).

[CON05] Bing-Yu Chen, Yutaka Ono et Tomoyuki Nishita. “Character animation creation using hand-drawn sketches”. In : *The Visual Computer* 21.8 (2005), p. 551-558 (cf. p. 43).

[Che+22] Qiang Chen, Tingsong Lu et al. “A Practical Model for Realistic Butterfly Flight Simulation”. In : *ACM Transactions on Graphics (TOG)* 41.3 (2022), p. 1-12 (cf. p. 41).

[Che+19] Yi Chen, Li Yu, Kaoru Ota et Mianxiong Dong. “Hierarchical Posture Representation for Robust Action Recognition”. In : *IEEE Trans. Comput. Soc. Syst.* 6.5 (2019), p. 1115-1125. url : https://doi.org/10.1109/TCSS.2019.2934639 (cf. p. 30, 35).

[Chi+00] Diane Chi, Monica Costa, Liwei Zhao et Norman Badler. “The EMOTE Model for Effort and Shape”. In : *Proceedings of the 27th Annual Conference on Computer Graphics and Interactive Techniques*. SIGGRAPH ’00. USA : ACM Press/Addison-Wesley Publishing Co., juill. 2000, p. 173-182 (cf. p. 108).

[Cla16] Simon Clavet. *Motion Matching for Realistic Animation in "For Honor"*. 2016 (cf. p. 45).

[Coh+03] Ira Cohen, Nicu Sebe, Ashutosh Garg, Lawrence S. Chen et Thomas S. Huang. “Facial Expression Recognition from Video Sequences : Temporal and Static Modeling”. In : *Computer Vision and Image Understanding*. Special Issue on Face Recognition 91.1 (juill. 2003), p. 160-187 (cf. p. 55).

[Con20] MMPose Contributors. *OpenMMLab Pose Estimation Toolbox and Benchmark*. https://github.com/open-mmlab/mmpose. 2020 (cf. p. 30, 35).

[CNK16] M. J. Cossetin, J. C. Nievola et A. L. Koerich. “Facial expression recognition using a pairwise feature selection and classification approach”. In : *2016 International Joint Conference on Neural Networks (IJCNN)*. 2016, p. 5149-5155 (cf. p. 23).

[CBM09] Matthieu Courgeon, Stéphanie Buisine et Jean-Claude Martin. "Impact of expressive wrinkles on perception of a virtual character's facial expressions of emotions". In : *International Workshop on Intelligent Virtual Agents*. Springer. 2009, p. 201-214 (cf. p. 22, 53, 88).

[CSB11] Roddy Cowie, Naomi Sussman et Aaron Ben-Ze'ev. "Emotion : Concepts and Definitions". In : *Emotion-Oriented Systems : The Humaine Handbook*. Sous la dir. de Roddy Cowie, Catherine Pelachaud et Paolo Petta. Cognitive Technologies. Berlin, Heidelberg : Springer, 2011, p. 9-30 (cf. p. 28).

[Cre19] Arthur Crenn. "Capture et transfert d'expression de visages d'enfants pour l'interaction avec des mondes virtuels". Theses. Université Claude Bernard Lyon 1, LIRIS, 2019 (cf. p. 12, 13, 22, 38, 61, 68, 70, 72, 81).

[Cre+16] Arthur Crenn, Rizwan Ahmed Khan, Alexandre Meyer et Saida Bouakaz. "Body expression recognition from animated 3D skeleton". In : (2016), p. 1-7 (cf. p. 65, 67, 68, 80, 81, 116).

[Cre+17a] Arthur Crenn, Hubert Konik, Alexandre Meyer et Saïda Bouakaz. "Reconnaissance d'expressions corporelles à l'aide d'un mouvement neutre synthétisé". In : *CORESA 2017 (COmpression et REpresentation des Signaux Audiovisuels)*. Laboratoire GREYC (UMR CNRS 6072). Caen, France, nov. 2017. url : https://hal.science/hal-01680819 (cf. p. 117).

[Cre+17b] Arthur Crenn, Alexandre Meyer, Rizwan Ahmed Khan, Hubert Konik et Saïda Bouakaz. "Toward an Efficient Body Expression Recognition Based on the Synthesis of a Neutral Movement". In : ICMI 2017 Proceedings of the 19th ACM International Conference on Multimodal Interaction (2017) (cf. p. 68, 80, 81, 116).

[Cre+20] Arthur Crenn, Alexandre Meyer, Hubert Konik, Rizwan Ahmed Khan et Saida Bouakaz. "Generic body expression recognition based on synthesis of realistic neutral motion". In : *IEEE Access* 8 (2020), p. 207758-207767 (cf. p. 68-72, 80, 81, 115).

[DGS13] Nele Dael, Martijn Goudbeek et Klaus R Scherer. "Perceived gesture dynamics in nonverbal expression of emotion". In : *Perception* 42.6 (2013), p. 642-657 (cf. p. 36).

[DF07] Sofia Dahl et Anders Friberg. "Visual Perception of Expressiveness in Musicians' Body Movements". In : *Music Perception : An Interdisciplinary Journal* 24.5 (juin 2007), p. 433-454 (cf. p. 79).

[DM14] M. Dahmane et J. Meunier. "Prototype-Based Modeling forFacial Expression Analysis". In : *IEEE Transactions on Multimedia* 16.6 (2014), p. 1574-1584 (cf. p. 23).

[DTS06] Navneet Dalal, Bill Triggs et Cordelia Schmid. "Human Detection Using Oriented Histograms of Flow and Appearance". In : *Computer Vision – ECCV 2006*. Sous la dir. d'Aleš Leonardis, Horst Bischof et Axel Pinz. Lecture Notes in Computer Science. Berlin, Heidelberg : Springer, 2006, p. 428-441 (cf. p. 23).

[DGD13] Kirsten A Dalrymple, Jesse Gomez et Brad Duchaine. “The Dartmouth Database of Children’s Faces : acquisition and validation of a new face stimulus set”. In : *PloS one* 8.11 (2013), e79131 (cf. p. 29).

[Dam02] Antonio R. Damasio. *Le Sentiment même de soi - Corps. émotions. conscience de soi*. Broché Poche, 2002 (cf. p. 17).

[Dao+17] Mohamed Daoudi, Stefano Berretti, Pietro Pala, Yvonne Delevoye et Alberto Del Bimbo. “Emotion recognition by body movement representation on the manifold of symmetric positive definite matrices”. In : *International Conference on Image Analysis and Processing*. Springer. 2017, p. 550-560 (cf. p. 37).

[Dav+06] James Davis, Maneesh Agrawala, Erika Chuang, Zoran Popović et David Salesin. “A Sketching Interface for Articulated Figure Animation”. In : *ACM SIGGRAPH 2006 Courses*. SIGGRAPH ’06. New York, NY, USA : Association for Computing Machinery, juill. 2006, 15-es (cf. p. 43).

[De +04] Berardina De Carolis, Catherine Pelachaud, Isabella Poggi et Mark Steedman. “APML, a Markup Language for Believable Behavior Generation”. In : *Life-Like Characters : Tools, Affective Functions, and Applications*. Sous la dir. d’Helmut Prendinger et Mitsuru Ishizuka. Cognitive Technologies. Berlin, Heidelberg : Springer, 2004, p. 65-85 (cf. p. 41).

[De +15] Fernando De la Torre, Wen-Sheng Chu et al. “IntraFace”. In : *2015 11th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*. T. 1. 2015, p. 1-8 (cf. p. 19).

[DB04] P. Ravindra De Silva et Nadia Bianchi-Berthouze. “Modeling human affective postures : an information theoretic characterization of posture features”. In : *Computer Animation and Virtual Worlds* 15.3-4 (2004), p. 269-276. eprint : https://onlinelibrary.wiley.com/doi/pdf/10.1002/cav.29. url : https://onlinelibrary.wiley.com/doi/abs/10.1002/cav.29 (cf. p. 33).

[Dem+14] Özlem Demir, Mehdi Labaied, Chris Merritt, Ken Stuart et Rommie E. Amaro. “Computer-Aided Discovery of Trypanosoma brucei RNA-Editing Terminal Uridylyl Transferase 2 Inhibitors”. In : *Chemical Biology and Drug Design* 84.2 (2014), p. 131-139 (cf. p. 24).

[Den+09] Jia Deng, Wei Dong et al. “Imagenet : A large-scale hierarchical image database”. In : *2009 IEEE conference on computer vision and pattern recognition*. Ieee. 2009, p. 248-255 (cf. p. 59).

[Des+21] Yann Desmarais, Denis Mottet, Pierre Slangen et Philippe Montesinos. “A review of 3D human pose estimation algorithms for markerless motion capture”. In : *Computer Vision and Image Understanding* 212 (2021), p. 103275 (cf. p. 30).

[DAS17] Swati Dewan, Shubham Agarwal et Navjyoti Singh. “Laban movement analysis to classify emotions from motion”. In : *Tenth International Conference on Machine Vision (ICMV 2017)*. International Society for Optics et Photonics, 2017, 106962Q (cf. p. 37, 80).

[Dha+12] A. Dhall, R. Goecke, S. Lucey et T. Gedeon. “Collecting Large, Richly Annotated Facial-Expression Databases from Movies”. In : *IEEE MultiMedia* 19.3 (2012), p. 34-41 (cf. p. 26).

[Dha+11] A. Dhall, R. Goecke, S. Lucey et T. Gedeon. “Static facial expression analysis in tough conditions : Data, evaluation protocol and benchmark”. In : *2011 IEEE International Conference on Computer Vision Workshops (ICCV Workshops)*. 2011, p. 2106-2112 (cf. p. 26).

[Dha+09] Abhinav Dhall, Lucey Tom Gedeon et al. *Acted Facial Expressions In The Wild Database*. 2009 (cf. p. 26).

[Dia+10] Abdunnaser Diaf, Riadh Ksantini, Boubakeur Boufama et Rachid Benlamri. “A novel human motion recognition method based on eigenspace”. In : *International Conference Image Analysis and Recognition*. Springer. 2010, p. 167-175 (cf. p. 46).

[Din+14] Yu Ding, Ken Prepin, Jing Huang, Catherine Pelachaud et Thierry Artières. “Laughter Animation Synthesis”. In : *Proceedings of the 2014 International Conference on Autonomous Agents and Multi-Agent Systems*. 2014, p. 773-780 (cf. p. 47).

[Dur+17] Funda Durupinar, Mubbasir Kapadia, Susan Deutsch, Michael Neff et Norman I. Badler. “Perform : Perceptual Approach for Adding Ocean Personality to Human Motion Using Laban Movement Analysis”. In : *ACM Transactions on Graphics (TOG)* 36.1 (2017), p. 6 (cf. p. 22).

[Dur+09] Funda Durupinar, Nuria Pelechano, Jan Allbeck, Uğur Güdükbay et Norman I Badler. “How the ocean personality model affects the perception of crowds”. In : *IEEE Computer Graphics and Applications* 31.3 (2009), p. 22-31 (cf. p. 21, 38).

[Dut11] Ludovic Dutreve. “Paramétrisation et Transfert d’Animations Faciales 3D à partir de Séquences Vidéo : vers des Applications en Temps Réel.” Theses. Université Claude Bernard Lyon 1, LIRIS, 2011 (cf. p. 12, 13, 87, 110).

[DMB11] Ludovic Dutreve, Alexandre Meyer et Saïda Bouakaz. “Easy acquisition and real-time animation of facial wrinkles”. In : *Computer Animation and Virtual Worlds* 22.2-3 (2011), p. 169-176 (cf. p. 109, 115).

[DMB08] Ludovic Dutreve, Alexandre Meyer et Saïda Bouakaz. “Feature points based facial animation retargeting”. In : *Proceedings of the 2008 ACM symposium on Virtual reality software and technology*. 2008, p. 197-200 (cf. p. 109, 116).

[Dut+10] Ludovic Dutreve, Alexandre Meyer, Veronica Orvalho et Saïda Bouakaz. “Easy rigging of face by automatic registration and transfer of skinning parameters”. In : *International Conference on Computer Vision and Graphics*. Springer. 2010, p. 333-341 (cf. p. 109, 116).

[Egg+11] Helen Link Egger, Daniel S Pine et al. “The NIMH Child Emotional Faces Picture Set (NIMH-ChEFS) : a new set of children’s facial emotion stimuli”. In : *International journal of methods in psychiatric research* 20.3 (2011), p. 145-156 (cf. p. 29).

[Ekm+87] P. Ekman, W. V. Friesen et al. “Universals and cultural differences in the judgments of facial expressions of emotion”. In : *J Pers Soc Psychol* 53.4 (1987), p. 712-717 (cf. p. 19).

[EF76] Paul Ekman et Wallace V Friesen. “Measuring facial movement”. In : *Environmental psychology and nonverbal behavior* 1.1 (1976), p. 56-75 (cf. p. 20).

[Eye] SR Research Eyelink II. *https ://www.sr-research.com/* (cf. p. 53).

[FSM16] C Fabian Benitez-Quiroz, Ramprakash Srinivasan et Aleix M Martinez. “Emotionet : An accurate, real-time algorithm for the automatic annotation of a million facial expressions in the wild”. In : *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016, p. 5562-5570 (cf. p. 29).

[Fil+19] Panagiotis Paraskevas Filntisis, Niki Efthymiou, Petros Koutras, Gerasimos Potamianos et Petros Maragos. “Fusing body posture with facial expressions for joint recognition of affect in child–robot interaction”. In : *IEEE Robotics and automation letters* 4.4 (2019), p. 4011-4018 (cf. p. 39).

[FP14] Nesrine Fourati et Catherine Pelachaud. “Emilya : Emotional body expression in daily actions database.” In : *LREC*. 2014, p. 3486-3493 (cf. p. 37, 38, 76, 78, 83).

[FP15] Nesrine Fourati et Catherine Pelachaud. “Multi-Level Classification of Emotional Body Expression”. In : *2015 11th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*. T. 1. Mai 2015, p. 1-8 (cf. p. 36).

[FP18] Nesrine Fourati et Catherine Pelachaud. “Perception of emotions and body movement in the Emilya database”. In : *IEEE Transactions on Affective Computing* (2018). url : `http://ieeexplore.ieee.org/document/7511699/` (visité le 1er déc. 2017) (cf. p. 70, 79).

[FPD19] Nesrine Fourati, Catherine Pelachaud et Patrice Darmon. “Contribution of Temporal and Multi-Level Body Cues to Emotion Classification”. In : *2019 8th International Conference on Affective Computing and Intelligent Interaction (ACII)*. Sept. 2019, p. 116-122 (cf. p. 36).

[Gaf+12] Yoren Gaffary, Victoria Eyharabide, Jean-Claude Martin et Mehdi Ammi. “Comparison of statistical methods for analysis of affective haptic expressions”. In : *Proceedings of the ACM Symposium on Applied Perception*. 2012, p. 128-128 (cf. p. 71).

[Gan+20] Yanling Gan, Jingying Chen, Zongkai Yang et Luhui Xu. “Multiple attention network for facial expression recognition”. In : *IEEE Access* 8 (2020), p. 7383-7393 (cf. p. 28).

[Gao+21] Pengcheng Gao, Ke Lu, Jian Xue, Jiayi Lyu et Ling Shao. “A facial landmark detection method based on deep knowledge transfer”. In : *IEEE Transactions on Neural Networks and Learning Systems* (2021) (cf. p. 89).

[GRC19] Maxime Garcia, Remi Ronfard et Marie-Paule Cani. “Spatial Motion Doodles : Sketching Animation in VR Using Hand Gestures and Laban Motion Analysis”. In : *Motion, Interaction and Games*. MIG ’19. New York, NY, USA : Association for Computing Machinery, oct. 2019, p. 1-10 (cf. p. 43).

[GEB16] Leon A Gatys, Alexander S Ecker et Matthias Bethge. “Image style transfer using convolutional neural networks”. In : *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016, p. 2414-2423 (cf. p. 48, 72, 103).

[GBK01] A.S. Georghiades, P.N. Belhumeur et D.J. Kriegman. “From Few to Many : Illumination Cone Models for Face Recognition under Variable Lighting and Pose”. In : *IEEE Trans. Pattern Anal. Mach. Intelligence* 23.6 (2001), p. 643-660 (cf. p. 29).

[GP03] M. A. Giese et T. Poggio. “Neural mechanisms for the recognition of biological movements”. In : *Nat. Rev. Neurosci.* 4.3 (2003), p. 179-192 (cf. p. 31).

[Gir+15] Tom Giraud, Florian Focone, Virginie Demulier, Jean-Claude Martin et Brice Isableu. “Perception of Emotion and Personality through Full-Body Movement Qualities : A Sport Coach Case Study”. In : *ACM Trans. Appl. Percept.* 13.1 (2015), 2 :1-2 :27. url : https://doi.org/10.1145/2791294 (cf. p. 33).

[Glo+11] Donald Glowinski, Nele Dael et al. “Toward a Minimal Representation of Affective Gestures”. In : *IEEE Transactions on Affective Computing* 2.2 (avr. 2011), p. 106-118 (cf. p. 79).

[Gro+04a] Keith Grochow, Steven L Martin, Aaron Hertzmann et Zoran Popović. “Style-based inverse kinematics”. In : *ACM SIGGRAPH 2004 Papers*. 2004, p. 522-531 (cf. p. 46).

[Gro+04b] Keith Grochow, Steven L. Martin, Aaron Hertzmann et Zoran Popović. “Style-based inverse kinematics”. In : *ACM Transactions on Graphics (TOG)*. T. 23. 00748. ACM, 2004, p. 522-531. url : http://dl.acm.org/citation.cfm?id=1015755 (visité le 23 oct. 2015) (cf. p. 46).

[GCF12] M. Melissa Gross, Elizabeth A. Crane et Barbara L. Fredrickson. “Effort-Shape and Kinematic Assessment of Bodily Expression of Emotion during Gait”. In : *Human Movement Science* 31.1 (fév. 2012), p. 202-221 (cf. p. 77, 79).

[GCF10] M. Melissa Gross, Elizabeth A. Crane et Barbara L. Fredrickson. “Methodology for Assessing Bodily Expression of Emotion”. In : *Journal of Nonverbal Behavior* 34.4 (déc. 2010), p. 223-248 (cf. p. 33, 36, 79).

[GCR13] Martin Guay, Marie-Paule Cani et Rémi Ronfard. “The Line of Action : an Intuitive Interface for Expressive Character Posing”. In : *ACM Transactions on Graphics* (nov. 2013) (cf. p. 43).

[Gua+15] Martin Guay, Rémi Ronfard, Michael Gleicher et Marie-Paule Cani. “Space-time sketching of character animation”. In : *ACM Transactions on Graphics*. Proceedings of ACM SIGGRAPH 2015 34.4 (août 2015), Article No. 118. url : https://hal.archives-ouvertes.fr/hal-01153763 (cf. p. 43, 107).

[GP06] Hatice Gunes et Massimo Piccardi. “Observer Annotation of Affective Display and Evaluation of Expressivity : Face vs. Face-and-Body”. In : *Proceedings of the HCSNet Workshop on Use of Vision in Human-Computer Interaction - Volume 56*. VisHCI ’06. Canberra, Australia : Australian Computer Society, Inc., nov. 2006, p. 35-42 (cf. p. 79).

[Guo+14] Shihui Guo, Jian Chang, Xiaosong Yang, Wencheng Wang et Jianjun Zhang. "Locomotion skills for insects with sample-based controller". In : *Computer Graphics Forum*. T. 33. 7. Wiley Online Library. 2014, p. 31-40 (cf. p. 40, 41).

[Hab+17] Ikhsanul Habibie, Daniel Holden, Jonathan Schwarz, Joe Yearsley et Taku Komura. "A recurrent variational autoencoder for human motion synthesis". In : *28th British Machine Vision Conference*. 2017 (cf. p. 75, 98).

[Hah+15] Fabian Hahn, Frederik Mutzel et al. "Sketch Abstractions for Character Posing". In : *Proceedings of the 14th ACM SIGGRAPH / Eurographics Symposium on Computer Animation*. SCA '15. New York, NY, USA : Association for Computing Machinery, août 2015, p. 185-191 (cf. p. 43).

[HBW15] Dennis Hamester, Pablo Barros et Stefan Wermter. "Face expression recognition with a 2-channel convolutional neural network". In : *2015 international joint conference on neural networks (IJCNN)*. IEEE. 2015, p. 1-8 (cf. p. 59).

[Han+13] Jungong Han, Ling Shao, Dong Xu et Jamie Shotton. "Enhanced computer vision with microsoft kinect sensor : A review". In : *IEEE transactions on cybernetics* 43.5 (2013), p. 1318-1334 (cf. p. 30).

[HR15] S. L. Happy et A. Routray. "Automatic facial expression recognition using features of salient facial patches". In : *IEEE Transactions on Affective Computing* 6.1 (2015), p. 1-12 (cf. p. 23).

[HR83] Jinni A. Harrigan et Robert Rosenthal. "Physicians' Head and Body Positions as Determinants of Perceived Rapport". In : *Journal of Applied Social Psychology* 13.6 (1983), p. 496-509. eprint : https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1559-1816.1983.tb02332.x. url : https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1559-1816.1983.tb02332.x (cf. p. 33).

[HSH16] G.P. Hegde, M. Seetha et Nagaratna Hegde. "Kernel Locality Preserving Symmetrical Weighted Fisher Discriminant Analysis based subspace approach for expression recognition". In : *Engineering Science and Technology, an International Journal* 19.3 (2016), p. 1321-1333. url : http://www.sciencedirect.com/science/article/pii/S2215098615300616 (cf. p. 22).

[Her+16] Andres Hernandez-Matamoros, Andrea Bonarini, Enrique Escamilla-Hernandez, Mariko Nakano-Miyatake et Hector Perez-Meana. "Facial expression recognition with automatic segmentation of face regions using a fuzzy based classification approach". In : *Knowledge-Based Systems* 110 (2016), p. 1-14. url : http://www.sciencedirect.com/science/article/pii/S0950705116302155 (cf. p. 22).

[HS99] John Hershberger et Subhash Suri. "An optimal algorithm for Euclidean shortest paths in the plane". In : *SIAM Journal on Computing* 28.6 (1999), p. 2215-2256 (cf. p. 94).

[HH06] M. Hirai et K. Hiraki. “The relative importance of spatial versus temporal structure in the perception of biological motion : an event-related potential study”. In : *Cognition* 99.1 (2006), p. 15-29 (cf. p. 31).

[Hol+17] Daniel Holden, Ikhsanul Habibie, Ikuo Kusajima et Taku Komura. “Fast Neural Style Transfer for Motion Data”. In : *IEEE Computer Graphics and Applications* 37.4 (2017), p. 42-49. url : http://ieeexplore.ieee.org/document/8013475/ (visité le 15 fév. 2018) (cf. p. 48, 97, 103).

[Hol+20a] Daniel Holden, Oussama Kanoun, Maksym Perepichka et Tiberiu Popa. “Learned Motion Matching”. In : *ACM Trans. Graph.* 39.4 (juill. 2020), 53 :53 :1-53 :53 :12 (cf. p. 45).

[Hol+20b] Daniel Holden, Oussama Kanoun, Maksym Perepichka et Tiberiu Popa. “Learned motion matching”. In : *ACM Transactions on Graphics (TOG)* 39.4 (2020), p. 53-1 (cf. p. 103).

[HKS17] Daniel Holden, Taku Komura et Jun Saito. “Phase-Functioned Neural Networks for Character Control”. In : *ACM Transactions on Graphics* 36.4 (juill. 2017), p. 1-13 (cf. p. 45, 97, 103, 107).

[HSK16] Daniel Holden, Jun Saito et Taku Komura. “A deep learning framework for character motion synthesis and editing”. In : *ACM Transactions on Graphics (TOG)* 35.4 (2016), p. 138 (cf. p. 45, 48, 73, 98, 103).

[Hol+15] Daniel Holden, Jun Saito, Taku Komura et Thomas Joyce. “Learning motion manifolds with convolutional autoencoders”. In : *SIGGRAPH Asia 2015 Technical Briefs*. ACM. 2015, p. 18 (cf. p. 48, 71, 73, 98).

[Hon+22] Fangzhou Hong, Mingyuan Zhang et al. “AvatarCLIP : Zero-Shot Text-Driven Generation and Animation of 3D Avatars”. In : *arXiv preprint arXiv :2205.08535* (2022) (cf. p. 113).

[HPP05] Eugene Hsu, Kari Pulli et Jovan Popović. “Style translation for human motion”. In : *ACM Transactions on Graphics (TOG)* 24.3 (2005). 00302, p. 1082-1089. url : http://dl.acm.org/citation.cfm?id=1073315 (visité le 23 oct. 2015) (cf. p. 46).

[Hu+22] Min Hu, Peng Ge, Xiaohua Wang, Hui Lin et Fuji Ren. “A spatio-temporal integrated model based on local and global features for video expression recognition”. In : *The Visual Computer* 38.8 (2022), p. 2617-2634 (cf. p. 25).

[Hua+17] Jing Huang, Qi Wang, Marco Fratarcangeli, Ke Yan et Catherine Pelachaud. “Multi-Variate Gaussian-Based Inverse Kinematics”. In : *Computer Graphics Forum* 36.8 (2017), p. 418-428 (cf. p. 102, 103).

[HV17] Zhiwu Huang et Luc Van Gool. “A riemannian network for spd matrix learning”. In : *Thirty-first AAAI conference on artificial intelligence*. 2017 (cf. p. 38).

[Jam84] William James. “What is an Emotion ?” In : *Mind* 9.34 (1884), p. 188-205. url : http://www.jstor.org/stable/2246769 (cf. p. 17).

[Jam32a] William T. James. “A Study of the Expression of Bodily Posture”. In : *The Journal of General Psychology* 7.2 (1932), p. 405-437. eprint : https://doi.org/10.1080/00221309.1932.9918475. url : https://doi.org/10.1080/00221309.1932.9918475 (cf. p. 33).

[Jam32b] William T. James. “A Study of the Expression of Bodily Posture”. In : *The Journal of General Psychology* 7.2 (oct. 1932), p. 405-437 (cf. p. 77, 79).

[Jin+19] Yongcheng Jing, Yezhou Yang et al. “Neural style transfer : A review”. In : *IEEE transactions on visualization and computer graphics* 26.11 (2019), p. 3365-3385 (cf. p. 19, 48, 103).

[Jol11] Ian Jolliffe. *Principal component analysis*. Springer, 2011 (cf. p. 71).

[Kac+18] Anis Kacem, Mohamed Daoudi, Boulbaba Ben Amor, Stefano Berretti et Juan Carlos Alvarez-Paiva. “A novel geometric framework on gram matrix trajectories for human behavior understanding”. In : *IEEE transactions on pattern analysis and machine intelligence* 42.1 (2018), p. 1-14 (cf. p. 37).

[Kal08] Marcelo Kallmann. “Analytical Inverse Kinematics with Body Posture Control”. In : *Computer animation and virtual worlds* 19.2 (2008), p. 79-91 (cf. p. 42).

[KLG97] Miyuki Kamachi, Michael Lyons et Jiro Gyoba. “The japanese female facial expression (jaffe) database”. In : *Availble : http ://www. kasrl. org/jaffe. html* (1997) (cf. p. 29).

[KCY00] T. Kanade, J. F. Cohn et Yingli Tian. “Comprehensive database for facial expression analysis”. In : *Proceedings Fourth IEEE International Conference on Automatic Face and Gesture Recognition (Cat. No. PR00580)*. 2000, p. 46-53 (cf. p. 29).

[KG16] Gu Eon Kang et M. Melissa Gross. “The Effect of Emotion on Movement Smoothness during Gait in Healthy Young Adults”. In : *Journal of Biomechanics* 49.16 (déc. 2016), p. 4022-4027 (cf. p. 79).

[Kap18a] A.I. Kapandji. *Anatomie Fonctionnelle. Volume 1. Membre Supérieur*. Seventh. T. 1. Maloine, 2018 (cf. p. 77).

[Kap18b] A.I. Kapandji. *Anatomie Fonctionnelle. Volume 2. Membre Inférieur*. Seventh. T. 2. Maloine, 2018 (cf. p. 77).

[Kap19] A.I. Kapandji. *Anatomie Fonctionnelle. Volume 3. Tête et Rachis*. Seventh. T. 3. Maloine, 2019 (cf. p. 77).

[KKB10] M. Karg, K. Kuhnlenz et M. Buss. “Recognition of Affect Based on Gait Patterns”. In : *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)* 40.4 (2010), p. 1050-1061 (cf. p. 36).

[Kar+11] Ahmad Abdul Karim, Thibaut Gaudin, Alexandre Meyer, Axel Buendia et Saïda Bouakaz. “Adding Physical Like Reaction Effects to Skeleton-Based Animations Using Controllable Pendulums”. In : *CASA 2011, Trans. Edutainment* 6 (2011), p. 111-121. url : https://doi.org/10.1007/978-3-642-22639-7%5C_12 (cf. p. 91-94, 110, 115).

[Kar+13] Ahmad Abdul Karim, Thibaut Gaudin, Alexandre Meyer, Axel Buendia et Saida Bouakaz. "Procedural Locomotion of Multilegged Characters in Dynamic Environments". In : *Computer Animation and Virtual Worlds* 24.1 (2013), p. 3-15 (cf. p. 94, 95, 109, 115).

[Kar+12] Ahmad Abdul Karim, Alexandre Meyer, Thibaut Gaudin, Axel Buendia et Saida Bouakaz. "Generic Spine Model with Simple Physics for Life-Like Quadrupeds and Reptiles." In : *VRIPHYS*. 2012, p. 97-106 (cf. p. 96, 109, 116).

[Kar+19] Sukumar Karumuri, Radoslaw Niewiadomski, Gualtiero Volpe et Antonio Camurri. "From motions to emotions : classification of affect from dance movements using deep learning". In : *Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems*. 2019, p. 1-6 (cf. p. 38).

[Kav+07] Ladislav Kavan, Steven Collins, Jiri Zara et Carol O'Sullivan. "Skinning with dual quaternions". In : *Proceedings of the 2007 symposium on Interactive 3D graphics and games*. 2007, p. 39-46 (cf. p. 88).

[Kha+20] R. A. Khan, A. Crenn, A. Meyer et S. Bouakaz. *LIRIS-CSE Children Spontaneous Facial Expression Video Database*. 2020. URL : `https://childrenfacialexpressi projet.liris.cnrs.fr` (cf. p. 56, 61, 117).

[Kha13] Rizwan A. Khan. "Expression recognition from videos in uncontrolled environment". Theses. Université Claude Bernard Lyon 1, LIRIS, 2013 (cf. p. 12, 13, 22, 52, 61).

[Kha+19a] Rizwan Ahmed Khan, Arthur Crenn, Alexandre Meyer et Saida Bouakaz. "A novel database of children's spontaneous facial expressions (LIRIS-CSE)". In : *Image and Vision Computing* 83-84 (2019), p. 61-69 (cf. p. 58-61, 115).

[KMB15] Rizwan Ahmed Khan, Alexandre Meyer et Saida Bouakaz. "Automatic affect analysis : from children to adults". In : *International Symposium on Visual Computing*. Springer. 2015, p. 304-313 (cf. p. 61, 116).

[Kha+13a] Rizwan Ahmed Khan, Alexandre Meyer, Hubert Konik et Saïda Bouakaz. "Pain detection through shape and appearance features". In : *Proceedings of the 2013 IEEE International Conference on Multimedia and Expo, ICME 2013, San Jose, CA, USA, July 15-19, 2013*. IEEE Computer Society, 2013, p. 1-6 (cf. p. 55, 61, 116).

[Kha+12a] Rizwan Ahmed Khan, Alexandre Meyer, Hubert Konik et Saida Bouakaz. "Exploring human visual system : study to aid the development of automatic facial expression recognition framework". In : *2012 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*. IEEE. 2012, p. 49-54 (cf. p. 60, 116).

[Kha+13b] Rizwan Ahmed Khan, Alexandre Meyer, Hubert Konik et Saida Bouakaz. "Framework for reliable, real-time facial expression recognition for low resolution images". In : *Pattern Recognition Letters* 34.10 (2013), p. 1159-1168 (cf. p. 52-55, 61, 115).

[Kha+12b] Rizwan Ahmed Khan, Alexandre Meyer, Hubert Konik et Saida Bouakaz. "Human Vision Inspired Framework for Facial Expressions Recognition". In : *2012 19th IEEE International Conference on Image Processing*. Sept. 2012, p. 2593-2596 (cf. p. 61, 116).

[Kha+19b] Rizwan Ahmed Khan, Alexandre Meyer, Hubert Konik et Saida Bouakaz. "Saliency-based framework for facial expression recognition". In : *Frontiers of Computer Science* 13.1 (2019), p. 183-198 (cf. p. 61, 115).

[Kim+20] SangBin Kim, Inbum Park, Seongsu Kwon et JungHyun Han. "Motion Retargetting Based on Dilated Convolutions and Skeleton-specific Loss Functions". In : *Computer Graphics Forum* 39.2 (2020), p. 497-507 (cf. p. 97).

[KM09] Michael Kipp et Jean-Claude Martin. "Gesture and emotion : Can basic gestural form features discriminate emotions ?" In : *2009 3rd International Conference on Affective Computing and Intelligent Interaction and Workshops*. 2009, p. 1-8 (cf. p. 36).

[KK81] Paul R. Kleinginna et Anne M. Kleinginna. "A categorized list of emotion definitions, with suggestions for a consensual definition". In : *Motivation and Emotion* 5.4 (1981), p. 345-379. url : https://doi.org/10.1007/BF00992553 (cf. p. 16).

[KBS11] A. Kleinsmith, N. Bianchi-Berthouze et A. Steed. "Automatic Recognition of Non-Acted Affective Postures". In : *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)* 41.4 (2011). 00125, p. 1027-1038. url : http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=5704207 (visité le 8 jan. 2016) (cf. p. 39, 77).

[KB13] Andrea Kleinsmith et Nadia Bianchi-Berthouze. "Affective body expression perception and recognition : A survey". In : *Affective Computing, IEEE Transactions on* 4.1 (2013). 00256, p. 15-33. url : http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=6212434 (visité le 20 nov. 2015) (cf. p. 35, 78).

[KB07] Andrea Kleinsmith et Nadia Bianchi-Berthouze. "Recognizing Affective Dimensions from Body Posture". In : *Proceedings of the 2Nd International Conference on Affective Computing and Intelligent Interaction*. ACII '07. Lisbon, Portugal : Springer-Verlag, 2007, p. 48-58. url : http://dx.doi.org/10.1007/978-3-540-74889-2_5 (cf. p. 33, 39).

[KDB06] Andrea Kleinsmith, P. Ravindra De Silva et Nadia Bianchi-Berthouze. "Cross-cultural differences in recognizing affect from body posture". In : *Interacting with Computers* 18.6 (2006). 00163, p. 1371-1389. url : http://www.sciencedirect.com/science/article/pii/S0953543806000634 (visité le 4 mai 2016) (cf. p. 33, 34, 38).

[KF22] Yu Kong et Yun Fu. "Human action recognition and prediction : A survey". In : *International Journal of Computer Vision* 130.5 (2022), p. 1366-1401 (cf. p. 30, 35, 37).

[KGP08] Lucas Kovar, Michael Gleicher et Frédéric Pighin. "Motion graphs". In : *ACM SIGGRAPH 2008 classes*. 2008, p. 1-10 (cf. p. 44).

[KPS13] Aung Sithu Kyaw, Clifford Peters et Thet Naing Swe. *Unity 4. x Game AI Programming*. Packt Publishing, 2013 (cf. p. 43).

[LU71] R. von Laban et L. Ullmann. *The mastery of movement*. The Mastery of Movement vol. 1971,ptie. 1. 00000 LCCN : b71023443. Macdonald et Evans, 1971. URL : https://books.google.fr/books?id=-RYIAQAAMAAJ (cf. p. 30, 31).

[LG15] Caroline Larboulette et Sylvie Gibet. "A Review of Computable Expressive Descriptors of Human Motion". In : *Proceedings of the 2nd International Workshop on Movement and Computing (MOCOŠ15)* (août 2015) (cf. p. 37, 71).

[Las01] John Lasseter. "Tricks to animating characters with a computer". In : *ACM Siggraph Computer Graphics* 35.2 (2001), p. 45-47 (cf. p. 43).

[LK05] Manfred Lau et James J Kuffner. "Behavior planning for character animation". In : *Proceedings of the 2005 ACM SIGGRAPH/Eurographics symposium on Computer animation*. 2005, p. 271-280 (cf. p. 43, 44).

[Law04] Neil D. Lawrence. "Gaussian process latent variable models for visualisation of high dimensional data". In : *Advances in neural information processing systems* 16.3 (2004). 00629, p. 329-336. URL : http://books.google.com/books?hl=en&lr=&id=0F-9C7K8fQ8C&oi=fnd&pg=PA329&dq=%22to+be+equivalent+to+that+solved+in+PCA+(see+e.g.+%5B10%5D),+indeed+the+formulation%22+%22Gaussian+Process+Latent+Variable%22+%22optimiser+such+as+scaled+conjugate+gradients+(SCG)+%5B7%5D+to+obtain+a+latent%22+&ots=THHvp-Ud82&sig=kkz7fFS2HZpkkIjsE4j9DAaMqDI (visité le 23 oct. 2015) (cf. p. 46).

[Laz20] Conor Lazarou. "Autoencoding generative adversarial networks". In : *arXiv preprint arXiv :2004.05472* (2020) (cf. p. 98).

[Le+12] Vuong Le, Jonathan Brandt, Zhe Lin, Lubomir Bourdev et Thomas S Huang. "Interactive facial feature localization". In : *European conference on computer vision*. Springer. 2012, p. 679-692 (cf. p. 23).

[LKF10] Yann LeCun, Koray Kavukcuoglu et Clément Farabet. "Convolutional networks and applications in vision". In : *Proceedings of 2010 IEEE International Symposium on Circuits and Systems*. IEEE. 2010, p. 253-256 (cf. p. 26, 59).

[LLL18] Kyungho Lee, Seyoung Lee et Jehee Lee. "Interactive character animation by learning multi-objective control". In : *ACM Transactions on Graphics (TOG)* 37.6 (2018), p. 1-10 (cf. p. 97).

[LB84] Stan Lee et John Buscema. *How to draw comics the Marvel way*. Simon et Schuster, 1984 (cf. p. 43).

[Lee+10] Yongjoon Lee, Kevin Wampler, Gilbert Bernstein, Jovan Popović et Zoran Popović. "Motion fields for interactive character locomotion". In : *ACM SIGGRAPH Asia 2010 papers*. 2010, p. 1-8 (cf. p. 44).

[LD20] Shan Li et Weihong Deng. "A deeper look at facial expression dataset bias". In : *IEEE Transactions on Affective Computing* (2020) (cf. p. 30, 113).

[LD22] Shan Li et Weihong Deng. “Deep Facial Expression Recognition : A Survey”. In : *IEEE Transactions on Affective Computing* 13.3 (2022), p. 1195-1215. URL : https://doi.org/10.1109%2Ftaffc.2020.2981446 (cf. p. 22, 23, 29, 37, 56).

[Li+13] Xiaobai Li, Tomas Pfister, Xiaohua Huang, Guoying Zhao et Matti Pietikäinen. “A spontaneous micro-expression database : Inducement, collection and baseline”. In : *2013 10th IEEE International Conference and Workshops on Automatic face and gesture recognition (fg)*. IEEE. 2013, p. 1-6 (cf. p. 58).

[Lia+22] Rijun Liao, Zhu Li, Shuvra S Bhattacharyya et George York. “PoseMapGait : A model-based gait recognition method with pose estimation maps and graph convolutional networks”. In : *Neurocomputing* 501 (2022), p. 514-528 (cf. p. 36, 113).

[Lim19] Jongin Lim. “PMnet : Learning of Disentangled Pose and Movement for Unsupervised Motion Retargeting”. In : *Proceedings of the 30th British Machine Vision Conference (BMVC 2019)*. Cardiff, United Kingdom : British Machine Vision Association, BMVA, sept. 2019, p. 14 (cf. p. 97).

[Liu+05] Ce Liu, Antonio Torralba, William T. Freeman, Frédo Durand et Edward H. Adelson. “Motion magnification”. In : *ACM transactions on graphics (TOG)* 24.3 (2005). 00182, p. 519-526. URL : http://dl.acm.org/citation.cfm?id=1073223 (visité le 25 jan. 2016) (cf. p. 46).

[Liu+21] Xin Liu, Henglin Shi et al. “iMiGUE : An identity-free video dataset for micro-gesture understanding and emotion analysis”. In : *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2021, p. 10631-10642 (cf. p. 38, 80).

[LKC21] Zhi-Song Liu, Vicky Kalogeiton et Marie-Paule Cani. “Multiple style transfer via variational autoencoder”. In : *2021 IEEE International Conference on Image Processing (ICIP)*. IEEE. 2021, p. 2413-2417 (cf. p. 75).

[Liu+20] Zhi-Song Liu, Wan-Chi Siu, Li-Wen Wang, Chu-Tak Li et Marie-Paule Cani. “Unsupervised real image super-resolution via generative variational autoencoder”. In : *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*. 2020, p. 442-443 (cf. p. 75).

[LT15] Vanessa LoBue et Cat Thrasher. “The Child Affective Facial Expression (CAFE) set : Validity and reliability from untrained adults”. In : *Frontiers in psychology* 5 (2015), p. 1532 (cf. p. 29).

[Luc+10] P. Lucey, J. F. Cohn et al. “The Extended Cohn-Kanade Dataset (CK+) : A complete dataset for action unit and emotion-specified expression”. In : *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition - Workshops*. 2010, p. 94-101 (cf. p. 29, 53).

[Luc+11] Patrick Lucey, Jeffrey F. Cohn, Kenneth M. Prkachin, Patricia E. Solomon et Iain Matthews. “Painful data : The UNBC-McMaster shoulder pain expression archive database”. In : *2011 IEEE International Conference on Automatic Face and Gesture Recognition (FG)*. 2011, p. 57-64 (cf. p. 55).

[Lug+19] Camillo Lugaresi, Jiuqiang Tang et al. *MediaPipe : A Framework for Building Perception Pipelines*. 2019. URL : https://arxiv.org/abs/1906.08172 (cf. p. 30, 35, 112).

[Ly+18] Son Thai Ly, Guee-Sang Lee, Soo-Hyung Kim et Hyung-Jeong Yang. "Emotion recognition via body gesture : Deep learning model coupled with keyframe selection". In : *Proceedings of the 2018 international conference on machine learning and machine intelligence*. 2018, p. 27-31 (cf. p. 38).

[Ma+10] Wanli Ma, Shihong Xia et al. "Modeling style and variation in human motion". In : *Proceedings of the 2010 ACM SIGGRAPH/Eurographics Symposium on Computer Animation*. 00035. Eurographics Association, 2010, p. 21-30. URL : http://dl.acm.org/citation.cfm?id=1921431 (visité le 23 oct. 2015) (cf. p. 46).

[MPP06] Yingliang Ma, Helena M. Paterson et Frank E. Pollick. "A motion capture library for the study of identity, gender, and emotion perception from biological motion". In : *Behavior research methods* (2006). 00130, p. 134-141. URL : http://link.springer.com/article/10.3758/BF03192758 (cf. p. 38).

[Mah23] Mehdi-Antoine Mahfoudi. "Génération procédurale d'animations porteuses d'expressions : approche temps réel et interactive". Theses. Université Claude Bernard Lyon 1, LIRIS, 2023 (cf. p. 13, 82, 110).

[Mah+22] Mehdi-Antoine Mahfoudi, Alexandre Meyer, Thibaut Gaudin, Axel Buendia et Saida Bouakaz. "Emotion Expression in Human Body Posture and Movement : A Survey on Intelligible Motion Factors, Quantification and Validation". In : *IEEE Transactions on Affective Computing* (2022), p. 1-24 (cf. p. 76-79, 81, 83, 84, 115).

[MAR98] A. M. MARTINEZ. "The AR face database". In : *CVC Technical Report24* (1998). URL : https://ci.nii.ac.jp/naid/10011462458/en/ (cf. p. 29).

[Mas+18] Ian Mason, Sebastian Starke, He Zhang, Hakan Bilen et Taku Komura. "Few-shot learning of homogeneous human locomotion styles". In : *Computer Graphics Forum*. T. 37. 7. Wiley Online Library. 2018, p. 143-153 (cf. p. 107).

[MKI09] Megumi Masuda, Shohei Kato et Hidenori Itoh. "Emotion Detection from Body Motion of Human Form Robot Based on Laban Movement Analysis". In : *Principles of Practice in Multi-Agent Systems*. Sous la dir. de Jung-Jin Yang, Makoto Yokoo, Takayuki Ito, Zhi Jin et Paul Scerri. Lecture Notes in Computer Science. Berlin, Heidelberg : Springer, 2009, p. 322-334 (cf. p. 79).

[McL96] P. McLeod. "Preserved and Impaired Detection of Structure From Motion by a 'Motion-blind" Patient". In : *Visual Cognition* 3.4 (1996), p. 363-392. eprint : https://doi.org/10.1080/135062896395634. URL : https://doi.org/10.1080/135062896395634 (cf. p. 31).

[Meh68] A. Mehrabian. "Inference of attitudes from the posture, orientation, and distance of a communicator". In : *J Consult Clin Psychol* 32.3 (1968), p. 296-308 (cf. p. 30, 33).

[Mei89] Marco de Meijer. “The contribution of general features of body movement to the attribution of emotions”. In : *Journal of Nonverbal Behavior* 13.4 (1989), p. 247-268. url : https://doi.org/10.1007/BF00990296 (cf. p. 33, 79).

[Men+19] Debin Meng, Xiaojiang Peng, Kai Wang et Yu Qiao. “Frame attention networks for facial expression recognition in videos”. In : *2019 IEEE international conference on image processing (ICIP)*. IEEE. 2019, p. 3866-3870 (cf. p. 28).

[MBB07] Alexandre Meyer, Hector M Briceno et Saida Bouakaz. “User-guided shape from shading to reconstruct fine details from a single photograph”. In : *Asian Conference on Computer Vision*. Springer. 2007, p. 738-747 (cf. p. 60, 90, 116).

[Mic+09] Johannes Michalak, Nikolaus F. Troje et al. “Embodiment of Sadness and Depression—Gait Patterns Associated With Dysphoric Mood”. In : *Psychosomatic Medicine* 71.5 (juin 2009), p. 580-587 (cf. p. 77, 79).

[MFM14] M.R. Mohammadi, E. Fatemizadeh et M.H. Mahoor. “PCA-based dictionary building for accurate facial expression recognition via sparse representation”. In : *Journal of Visual Communication and Image Representation* 25.5 (2014), p. 1082-1092. url : https://www.sciencedirect.com/science/article/pii/S1047320314000625 (cf. p. 25).

[Moh+09] David Moher, Alessandro Liberati, Jennifer Tetzlaff et Douglas G. Altman. “Preferred Reporting Items for Systematic Reviews and Meta-Analyses : The PRISMA Statement”. In : *Annals of Internal Medicine* 151.4 (août 2009), p. 264-269 (cf. p. 76).

[MHM17] Ali Mollahosseini, Behzad Hasani et Mohammad H Mahoor. “Affectnet : A database for facial expression, valence, and arousal computing in the wild”. In : *IEEE Transactions on Affective Computing* 10.1 (2017), p. 18-31 (cf. p. 29).

[Mon18] Lorenza Mondada. “Multiple temporalities of language and body in interaction : Challenges for transcribing multimodality”. In : *Research on Language and Social Interaction* 51.1 (2018), p. 85-106 (cf. p. 30).

[Mon+99] Joann Montepare, Elissa Koff, Deborah Zaitchik et Marilyn Albert. “The Use of Body Movements and Gestures as Cues to Emotions in Younger and Older Adults”. In : *Journal of Nonverbal Behavior* 23.2 (juin 1999), p. 133-152 (cf. p. 79).

[MGC87] Joann M. Montepare, Sabra B. Goldstein et Annmarie Clausen. “The Identification of Emotions from Gait Information”. In : *Journal of Nonverbal Behavior* 11.1 (mars 1987), p. 33-42 (cf. p. 77, 79).

[MK09] Tomohiko Mukai et Shigeru Kuriyama. “Pose-timeline for propagating motion edits”. In : *Proceedings of the 2009 acm siggraph/eurographics symposium on computer animation*. 2009, p. 113-122 (cf. p. 106, 107).

[Mul+99] Franck Multon, Laure France, Marie-Paule Cani-Gascuel et Giles Debunne. “Computer animation of human walking : a survey”. In : *The journal of visualization and computer animation* 10.1 (1999), p. 39-54 (cf. p. 41).

[Muy12] Eadweard Muybridge. *Animals in motion*. Courier Corporation, 2012 (cf. p. 95).

[NLG20] Lucie Naert, Caroline Larboulette et Sylvie Gibet. "A survey on the animation of signing avatars : From sign representation to utterance synthesis". In : *Computers & Graphics* 92 (2020), p. 76-98 (cf. p. 40).

[NH10] Vinod Nair et Geoffrey E Hinton. "Rectified linear units improve restricted boltzmann machines". In : *Icml*. 2010 (cf. p. 98).

[NAK02] T. NAKATA. "Analysis of Impression of Robot Bodily Expression". In : *Journal of Robotics and Mechatronics* 14.1 (2002), p. 27-36 (cf. p. 79).

[Nef+10] Michael Neff, Yingying Wang, Rob Abbott et Marilyn Walker. "Evaluating the effect of gesture and language on personality perception in conversational agents". In : *International Conference on Intelligent Virtual Agents*. Springer. 2010, p. 222-235 (cf. p. 21).

[Nes15] Fourati Nesrine. "Classification and Characterization of Emotional Body Expression in Daily Actions". Thèse de doct. Telecom Paristech, 2015 (cf. p. 80).

[Ng+15] Hong-Wei Ng, Viet Dung Nguyen, Vassilios Vonikakis et Stefan Winkler. "Deep learning for emotion recognition on small datasets using transfer learning". In : *Proceedings of the 2015 ACM on international conference on multimodal interaction*. 2015, p. 443-449 (cf. p. 27, 59).

[NSM18] Swati Nigam, Rajiv Singh et AK Misra. "Efficient facial expression recognition using histogram of oriented gradients in wavelet domain". In : *Multimedia tools and applications* 77.21 (2018), p. 28725-28747 (cf. p. 23).

[Nor+13] Aline Normoyle, Fannie Liu, Mubbasir Kapadia, Norman I. Badler et Sophie Jörg. "The effect of posture and dynamics on the perception of emotion". In : *Proceedings of the ACM Symposium on Applied Perception*. 00017. ACM, 2013, p. 91-98. URL : http://dl.acm.org/citation.cfm?id=2492500 (visité le 25 jan. 2016) (cf. p. 33, 34).

[Nor+18] Fatemeh Noroozi, Ciprian Adrian Corneanu et al. "Survey on emotional body gesture recognition". In : *IEEE transactions on affective computing* 12.2 (2018), p. 505-523 (cf. p. 30, 35-37, 80).

[OPM02] Timo Ojala, Matti Pietikainen et Topi Maenpaa. "Multiresolution gray-scale and rotation invariant texture classification with local binary patterns". In : *IEEE Transactions on pattern analysis and machine intelligence* 24.7 (2002), p. 971-987 (cf. p. 54).

[OG07a] Lars Omlor et Martin A. Giese. "Extraction of Spatio-Temporal Primitives of Emotional Body Expressions". In : *Neurocomputing*. Computational Neuroscience : Trends in Research 2007 70.10 (juin 2007), p. 1938-1942 (cf. p. 77).

[OG07b] Lars Omlor et Martin A. Giese. "Extraction of spatio-temporal primitives of emotional body expressions". In : *Neurocomputing* 70.10 (2007). 00034, p. 1938-1942. URL : http://www.sciencedirect.com/science/article/pii/S0925231206004309 (visité le 9 mai 2017) (cf. p. 31, 36, 77).

[OZM14] Ebenezer Owusu, Yongzhao Zhan et Qi Rong Mao. "A neural-AdaBoost based facial expression recognition system". In : *Expert Systems with Applications* 41.7 (2014), p. 3383-3390. URL : http://www.sciencedirect.com/science/article/pii/S0957417413009615 (cf. p. 22).

[Özt+13] A Cengiz Öztireli, Ilya Baran et al. "Differential blending for expressive sketch-based posing". In : *Proceedings of the 12th ACM SIGGRAPH/Eurographics Symposium on Computer Animation*. 2013, p. 155-164 (cf. p. 43).

[PY09] Sinno Jialin Pan et Qiang Yang. "A survey on transfer learning". In : *IEEE Transactions on knowledge and data engineering* 22.10 (2009), p. 1345-1359 (cf. p. 27, 59).

[PBD22] Giancarlo Paoletti, Cigdem Beyan et Alessio Del Bue. "Graph Laplacian-Improved Convolutional Residual Autoencoder for Unsupervised Human Action and Emotion Recognition". In : *IEEE Access* 10 (2022), p. 131128-131143 (cf. p. 80).

[Par+04] Hanhoon Park, Jong-Il Park, Un-Mi Kim et Woontack Woo. "Emotion Recognition from Dance Image Sequences Using Contour Approximation". In : *Structural, Syntactic, and Statistical Pattern Recognition*. Sous la dir. d'Ana Fred, Terry M. Caelli, Robert P. W. Duin, Aurélio C. Campilho et Dick de Ridder. Berlin, Heidelberg : Springer Berlin Heidelberg, 2004, p. 547-555 (cf. p. 36).

[PSS02] Sang Il Park, Hyun Joon Shin et Sung Yong Shin. "On-line locomotion generation based on motion blending". In : *Proceedings of the 2002 ACM SIGGRAPH/Eurographics symposium on Computer animation*. 2002, p. 105-111 (cf. p. 44).

[PWD06] M. V. Peelen, A. J. Wiggett et P. E. Downing. "Patterns of fMRI activity dissociate overlapping functional brain areas that respond to biological motion". In : *Neuron* 49.6 (2006), p. 815-822 (cf. p. 31).

[Pen+18] Xue Bin Peng, Pieter Abbeel, Sergey Levine et Michiel Van de Panne. "Deepmimic : Example-guided deep reinforcement learning of physics-based character skills". In : *ACM Transactions On Graphics (TOG)* 37.4 (2018), p. 1-14 (cf. p. 45).

[Pen+17] Xue Bin Peng, Glen Berseth, KangKang Yin et Michiel Van De Panne. "Deeploco : Dynamic locomotion skills using hierarchical deep reinforcement learning". In : *ACM Transactions on Graphics (TOG)* 36.4 (2017), p. 1-13 (cf. p. 45).

[Pia+13] Stefano Piana, Alessandra Stagliano, Antonio Camurri et Francesca Odone. "A Set of Full-Body Movement Features for Emotion Recognition to Help Children Affected by Autism Spectrum Condition". In : *IDGEI International Workshop*. 2013 (cf. p. 79).

[Pia+16] Stefano Piana, Alessandra Staglianò, Francesca Odone et Antonio Camurri. "Adaptive body gesture representation for automatic emotion recognition". In : *ACM Transactions on Interactive Intelligent Systems (TiiS)* 6.1 (2016), p. 1-31 (cf. p. 36).

[Plu80] Robert Plutchik. “A general psychoevolutionary theory of emotion”. In : *Theories of emotion*. Elsevier, 1980, p. 3-33 (cf. p. 21).

[Pol+01] Frank E Pollick, Helena M Paterson, Armin Bruderlin et Anthony J Sanford. “Perceiving Affect from Arm Movement”. In : *Cognition* 82.2 (déc. 2001), B51-B61 (cf. p. 79).

[PF06] Emmanuel Prados et Olivier Faugeras. “Shape from shading”. In : *Handbook of mathematical models in computer vision*. Springer, 2006, p. 375-388 (cf. p. 90).

[PL16] Liliana Lo Presti et Marco La Cascia. “3D skeleton-based human action classification : A survey”. In : *Pattern Recognition* 53 (2016), p. 130-147 (cf. p. 35).

[Qin+22] Xiaofei Qin, Rui Cai, Jiabin Yu, Changxiang He et Xuedian Zhang. “An efficient self-attention network for skeleton-based action recognition”. In : *Scientific Reports* 12.1 (2022), p. 1-10 (cf. p. 38, 81, 112).

[RR93] Madhusudan Raghavan et Bernard Roth. “Inverse kinematics of the general 6R manipulator and related linkages”. In : (1993) (cf. p. 42).

[Ram+21a] Aditya Ramesh, Mikhail Pavlov et al. “Zero-shot text-to-image generation”. In : *International Conference on Machine Learning*. PMLR. 2021, p. 8821-8831 (cf. p. 75).

[Ram+21b] Aditya Ramesh, Mikhail Pavlov et al. “Zero-shot text-to-image generation”. In : *International Conference on Machine Learning*. PMLR. 2021, p. 8821-8831 (cf. p. 113).

[Ran+19] Tanmay Randhavane, Aniket Bera et al. “EVA : Generating Emotional Behavior of Virtual Agents using Expressive Features of Gait and Gaze”. In : *ACM Symposium on Applied Perception 2019, SAP 2018, Barcelona, Spain, September 19-20, 2019*. Sous la dir. de Solene Neyret, Elena Kokkinara et al. ACM, 2019, 6 :1-6 :10. url : https://doi.org/10.1145/3343036.3343129 (cf. p. 48).

[Ran+20] Tanmay Randhavane, Uttaran Bhattacharya et al. “Identifying Emotions from Walking Using Affective and Deep Features”. In : *arXiv :1906.11884 [cs]* (jan. 2020). arXiv : 1906.11884 [cs] (cf. p. 38).

[Ran+22] Tanmay Randhavane, Uttaran Bhattacharya et al. “Learning Gait Emotions Using Affective and Deep Features”. In : *Proceedings of the ACM SIGGRAPH Conference on Motion, Interaction and Games*. 2022, p. 1-10 (cf. p. 36, 48).

[RWV19] Sarah Ribet, Hazem Wannous et Jean-Philippe Vandeborre. “Survey on style in 3d human body motion : Taxonomy, data, recognition and its applications”. In : *IEEE Transactions on Affective Computing* 12.4 (2019), p. 928-948 (cf. p. 35).

[Roe+09a] C. L. Roether, L. Omlor, A. Christensen et M. A. Giese. “Critical features for the perception of emotion from gait”. In : *Journal of Vision* 9.6 (2009), p. 1-32 (cf. p. 36).

[Roe+09b] Claire L. Roether, Lars Omlor, Andrea Christensen et Martin A. Giese. "Critical Features for the Perception of Emotion from Gait". In : *Journal of Vision* 9.6 (juin 2009), p. 15-15 (cf. p. 77).

[Rus03] J. A. Russell. "Core affect and the psychological construction of emotion". In : *Psychol Rev* 110.1 (2003), p. 145-172 (cf. p. 34).

[Rus80] James A Russell. "A circumplex model of affect." In : *Journal of personality and social psychology* 39.6 (déc. 1980), p. 1161 (cf. p. 20).

[SH07] Alla Safonova et Jessica K Hodgins. "Construction and optimal search of interpolated motion graphs". In : *ACM SIGGRAPH 2007 papers*. 2007, 106-es (cf. p. 44).

[SS18] Hanan Salam et Renaud Seguier. "A survey on face modeling : building a bridge between face analysis and synthesis". In : *The Visual Computer* 34.2 (2018), p. 289-319 (cf. p. 88).

[Sam+13] Ali-Akbar Samadani, Sarahjane Burton, Rob Gorbet et Dana Kulic. "Laban Effort and Shape Analysis of Affective Hand and Arm Movements". In : *2013 Humaine Association Conference on Affective Computing and Intelligent Interaction*. Sept. 2013, p. 343-348 (cf. p. 79).

[SI92] Ashok Samal et Prasana A Iyengar. "Automatic recognition and analysis of human faces and facial expressions : A survey". In : *Pattern recognition* 25.1 (1992), p. 65-77 (cf. p. 51).

[San+17] Anush Sankaran, Mayank Vatsa, Richa Singh et Angshul Majumdar. "Group sparse autoencoder". In : *Image and Vision Computing* 60 (2017), p. 64-74 (cf. p. 74).

[SG19] R. Santhoshkumar et M. Kalaiselvi Geetha. "Deep Learning Approach for Emotion Recognition from Human Body Movements with Feedforward Deep Convolution Neural Networks". In : *Procedia Computer Science* 152 (2019). International Conference on Pervasive Computing Advances and Applications-PerCAA 2019, p. 158-165. URL : https://www.sciencedirect.com/science/article/pii/S1877050919306908 (cf. p. 37).

[Sap+18] Tomasz Sapinski, Dorota Kaminska et al. "Multimodal Database of Emotional Speech, Video and Gestures". In : *Pattern Recognition and Information Forensics - ICPR 2018 International Workshops, CVAUI, IWCF, and MIPPSNA, Beijing, China, August 20-24, 2018, Revised Selected Papers*. 2018, p. 153-163 (cf. p. 38).

[SF18] Disa A Sauter et Agneta H Fischer. "Can perceivers recognise emotions from spontaneous expressions ?" In : *Cognition and Emotion* 32.3 (2018), p. 504-515 (cf. p. 28).

[SM08] Yann Savoye et Alexandre Meyer. "Multi-Layer Level of Detail For Character Animation". In : *Proceedings of the Fifth Workshop on Virtual Reality Interactions and Physical Simulations, VRIPHYS 2008, Grenoble, France, 2008*. Sous la dir. de François Faure et Matthias Teschner. Eurographics Association, 2008, p. 57-66. URL : https://doi.org/10.2312/PE/vriphys/vriphys08/057-066 (cf. p. 109, 116).

[SSI03] Misako Sawada, Kazuhiro Suda et Motonobu Ishii. “Expression of Emotions in Dance : Relation between Arm Movement Characteristics and Emotion”. In : *Perceptual and Motor Skills* 97.3 (déc. 2003), p. 697-708 (cf. p. 79).

[Sch+04] Klaus R. Scherer, Tanja Wranik, Janique Sangsue, Véronique Tran et Ursula Scherer. “Emotions in everyday life : probability of occurrence, risk factors, appraisal and reaction patterns”. In : *Social Science Information* 43.4 (2004), p. 499-570 (cf. p. 17).

[SVD08] Konrad Schindler, Luc Van Gool et Beatrice De Gelder. “Recognizing emotions expressed by body pose : A biologically inspired neural model”. In : *Neural networks* 21.9 (2008), p. 1238-1246 (cf. p. 31).

[Sen+16] Simon Senecal, Louis Cuel, Andreas Aristidou et Nadia Magnenat-Thalmann. “Continuous body emotion recognition system during theater performances”. In : *Computer Animation and Virtual Worlds* 27.3-4 (2016), p. 311-320 (cf. p. 36).

[ST11] Azim F Shariff et Jessica L Tracy. “What are emotion expressions for ?” In : *Current Directions in Psychological Science* 20.6 (2011), p. 395-399 (cf. p. 16).

[She+19] Zhijuan Shen, Jun Cheng, Xiping Hu et Qian Dong. “Emotion recognition based on multi-view body gestures”. In : *2019 IEEE International Conference on Image Processing (ICIP)*. IEEE. 2019, p. 3317-3321 (cf. p. 35).

[Shi+22] Henglin Shi, Wei Peng, Haoyu Chen, Xin Liu et Guoying Zhao. “Multiscale 3D-Shift Graph Convolution Network for Emotion Recognition From Human Actions”. In : *IEEE Intelligent Systems* 37.4 (2022), p. 103-110 (cf. p. 38, 80).

[Sid+15] M. H. Siddiqi, R. Ali, A. M. Khan, Y. Park et S. Lee. “Human Facial Expression Recognition Using Stepwise Linear Discriminant Analysis and Hidden Conditional Random Fields”. In : *IEEE Transactions on Image Processing* 24.4 (2015), p. 1386-1398 (cf. p. 25).

[SZ14] Karen Simonyan et Andrew Zisserman. “Very deep convolutional networks for large-scale image recognition”. In : *arXiv preprint arXiv :1409.1556* (2014) (cf. p. 59, 72).

[Smi+19] Harrison Jesse Smith, Chen Cao, Michael Neff et Yingying Wang. “Efficient neural networks for real-time motion style transfer”. In : *Proceedings of the ACM on Computer Graphics and Interactive Techniques* 2.2 (2019), p. 1-17 (cf. p. 48).

[SBW17] Sima Soltanpour, Boubakeur Boufama et QM Jonathan Wu. “A survey of local feature methods for 3D face recognition”. In : *Pattern Recognition* 72 (2017), p. 391-406 (cf. p. 22).

[Son+10] M. Song, D. Tao, Z. Liu, X. Li et M. Zhou. “Image Ratio Features for Facial Expression Recognition Application”. In : *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)* 40.3 (2010), p. 779-788 (cf. p. 23).

[Spi22] SpirOps. *Animation Procédurale*. 2022. url : `https://animation.spirops.com/?lg=fr` (cf. p. 107).

[Sta+19] Sebastian Starke, He Zhang, Taku Komura et Jun Saito. “Neural state machine for character-scene interactions.” In : *ACM Trans. Graph.* 38.6 (2019), p. 209-1 (cf. p. 45).

[Sta+20] Sebastian Starke, Yiwei Zhao, Taku Komura et Kazi Zaman. “Local motion phases for learning multi-contact character movements”. In : *ACM Transactions on Graphics (TOG)* 39.4 (2020), p. 54-1 (cf. p. 45).

[Ste+17] Benjamin Stephens-Fripp, Fazel Naghdy, David Stirling et Golshah Naghdy. “Automatic affect perception based on body gait and posture : A survey”. In : *International Journal of Social Robotics* 9.5 (2017), p. 617-641 (cf. p. 35).

[Tay14] Fredric W. Taylor. “Epilogue”. In : *The Scientific Exploration of Venus*. Cambridge University Press, 2014, p. 277-278 (cf. p. 25).

[TJT95] Frank Thomas, Ollie Johnston et Frank Thomas. *The illusion of life : Disney animation*. Hyperion New York, 1995 (cf. p. 30).

[TBP04] Matthew Thorne, David Burke et Michiel van de Panne. “Motion doodles : an interface for sketching character motion”. In : *ACM Transactions on Graphics (TOG)* 23.3 (2004), p. 424-431 (cf. p. 43).

[TD77] Patrick Tort et Charles Darwin. “L’Expression des émotions : Chez l’Homme et les animaux. Traduction et édition. Précède de L’origine de la sympathie de Patrick Tort”. In : *L’Expression des émotions* (1877), p. 1-672 (cf. p. 20).

[Tov+21] Omer Tov, Yuval Alaluf, Yotam Nitzan, Or Patashnik et Daniel Cohen-Or. “Designing an encoder for stylegan image manipulation”. In : *ACM Transactions on Graphics (TOG)* 40.4 (2021), p. 1-14 (cf. p. 48).

[TBZ16] Arthur Truong, Hugo Boujut et Titus Zaharia. “Laban descriptors for gesture recognition and emotional analysis”. en. In : *The Visual Computer* 32.1 (2016). 00000, p. 83-98. url : http://link.springer.com/10.1007/s00371-014-1057-8 (visité le 15 fév. 2016) (cf. p. 37, 38, 80).

[TBL18] Michael Tschannen, Olivier Frederic Bachem et Mario Lučić. “Recent Advances in Autoencoder-Based Representation Learning”. In : *Bayesian Deep Learning Workshop, NeurIPS*. 2018 (cf. p. 73).

[TV22] Arte TV. *Autopsie d’une Intelligence artificielle*. 2022. url : https://www.youtube.com/watch?v=QQkleleBnZQ (cf. p. 22).

[UAT95] Munetoshi Unuma, Ken Anjyo et Ryozo Takeuchi. “Fourier principles for emotion-based human figure animation”. In : *Proceedings of the 22nd annual conference on Computer graphics and interactive techniques*. 00582. ACM, 1995, p. 91-96. url : http://dl.acm.org/citation.cfm?id=218419 (visité le 17 nov. 2015) (cf. p. 47).

[Vai+90] L. M. Vaina, M. Lemay, D. C. Bienfang, A. Y. Choi et K. Nakayama. “Intact "biological motion" and "structure from motion" perception in a patient with impaired motion mechanisms : a case study”. In : *Vis. Neurosci.* 5.4 (1990), p. 353-369 (cf. p. 31).

[VP10] Michel Valstar et Maja Pantic. “Induced disgust, happiness and surprise : an addition to the mmi facial expression database”. In : *Proc. 3rd Intern. Workshop on EMOTION (satellite of LREC) : Corpora for Research on Emotion and Affect*. Paris, France. 2010, p. 65 (cf. p. 29, 56, 57).

[Vas+17] Ashish Vaswani, Noam Shazeer et al. “Attention is all you need”. In : *Advances in neural information processing systems* 30 (2017) (cf. p. 38).

[Vic23] Léon Victor. “Learning-Based Interactive Character Animation : Expressing Emotions through Motion”. Theses. INSA de Lyon, LIRIS, 2023 (cf. p. 12, 13, 99, 103, 110).

[VM20] Léon Victor et Alexandre Meyer. “Character Pose Design in Latent Space For Animation Edition”. In : *Journées Françaises de l’Informatique Graphique 2020*. Nancy, France, nov. 2020. url : https://hal.science/hal-03338910 (cf. p. 117).

[VMB21] Léon Victor, Alexandre Meyer et Saïda Bouakaz. “Learning-based pose edition for efficient and interactive design”. In : *Computer Animation and Virtual Worlds* 32.3-4 (2021), e2013 (cf. p. 100-102, 110, 115).

[VMB23] Léon Victor, Alexandre Meyer et Saïda Bouakaz. “Pose Metrics : a New Paradigm for Character Motion Edition”. In : *CoRR* abs/2301.06514 (2023). url : https://doi.org/10.48550/arXiv.2301.06514 (cf. p. 117).

[Vil+18] Ruben Villegas, Jimei Yang, Duygu Ceylan et Honglak Lee. “Neural Kinematic Networks for Unsupervised Motion Retargetting”. In : *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*. Juin 2018, p. 8639-8648 (cf. p. 97).

[VBP20] Vera Vine, Ryan L Boyd et James W Pennebaker. “Natural emotion vocabularies as windows on distress and well-being”. In : *Nature communications* 11.1 (2020), p. 1-9 (cf. p. 112).

[Vol+14] Ekaterina Volkova, Stephan de la Rosa, Heinrich H. Bülthoff et Betty Mohler. “The MPI Emotional Body Expressions Database for Narrative Scenarios”. en. In : *PLoS ONE* 9.12 (2014). Sous la dir. de Marko Nardini. 00000, e113647. url : http://dx.plos.org/10.1371/journal.pone.0113647 (visité le 26 avr. 2017) (cf. p. 36, 38).

[VCH94] Rudolf Von Laban, Jacqueline Challet-Haas et Marion Hansen. *La maîtrise du mouvement*. Actes sud, 1994 (cf. p. 30, 31).

[Wal98] Harald G. Wallbott. “Bodily expression of emotion”. In : *European Journal of Social Psychology* 28.6 (1998), p. 879-896 (cf. p. 33, 34, 77, 79).

[WS86] Harald G. Wallbott et Klaus R. Scherer. “Cues and Channels in Emotion Recognition”. In : *Journal of Personality and Social Psychology* 51.4 (1986), p. 690-699 (cf. p. 79).

[Wal+06] Frank Wallhoff, Björn Schuller, Michael Hawellek et Gerhard Rigoll. "Efficient Recognition of Authentic Dynamic Facial Expressions on the Feedtum Database." In : *ICME*. IEEE Computer Society, 2006, p. 493-496. URL : http://dblp.uni-trier.de/db/conf/icmcs/icme2006.html#WallhoffSHR06 (cf. p. 55).

[WC91a] Li-Chun Tommy Wang et Chih Cheng Chen. "A Combined Optimization Method for Solving the Inverse Kinematics Problems of Mechanical Manipulators". In : *IEEE Trans. Robot. Automat.* 7.4 (1991), p. 489-499 (cf. p. 103).

[WFH07] Jack M. Wang, David J. Fleet et Aaron Hertzmann. "Multifactor Gaussian process models for style-content separation". In : *Proceedings of the 24th international conference on Machine learning*. 00000. ACM, 2007, p. 975-982. URL : http://dl.acm.org/citation.cfm?id=1273619 (visité le 19 nov. 2015) (cf. p. 46).

[WFH09] Jack M. Wang, David J. Fleet et Aaron Hertzmann. "Optimizing Walking Controllers". In : *ACM SIGGRAPH Asia 2009 Papers*. SIGGRAPH Asia '09. Yokohama, Japan : Association for Computing Machinery, déc. 2009, p. 1-8 (cf. p. 41).

[WC91b] L-CT Wang et Chih-Cheng Chen. "A combined optimization method for solving the inverse kinematics problems of mechanical manipulators". In : *IEEE Transactions on Robotics and Automation* 7.4 (1991), p. 489-499 (cf. p. 42).

[WHK20] Lei Wang, Du Q. Huynh et Piotr Koniusz. "A Comparative Review of Recent Kinect-Based Action Recognition Algorithms". In : *IEEE Transactions on Image Processing* 29 (2020), p. 15-28 (cf. p. 81, 112).

[Wan+18a] Nannan Wang, Xinbo Gao, Dacheng Tao, Heng Yang et Xuelong Li. "Facial feature point detection : A comprehensive survey". In : *Neurocomputing* 275 (2018), p. 50-65 (cf. p. 24).

[Wan+20a] Qi Wang, Thierry Artières, Mickael Chen et Ludovic Denoyer. "Adversarial learning for modeling human motion". In : *The Visual Computer* 36.1 (2020), p. 141-160 (cf. p. 48).

[Wan+18b] Qi Wang, Mickaël Chen, Thierry Artières et Ludovic Denoyer. "Transferring style in motion capture sequences with adversarial learning". In : *ESANN*. 2018 (cf. p. 48).

[Wan+22a] Yan Wang, Wei Song et al. "A systematic review on affective computing : Emotion models, databases, and recent advances". In : *Information Fusion* (2022) (cf. p. 35).

[Wan+22b] Yan Wang, Yixuan Sun et al. "Ferv39k : a large-scale multi-scene dataset for facial expression recognition in videos". In : *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2022, p. 20922-20931 (cf. p. 29).

[Wan+20b] Yaqing Wang, Quanming Yao, James T Kwok et Lionel M Ni. "Generalizing from a few examples : A survey on few-shot learning". In : *ACM computing surveys (csur)* 53.3 (2020), p. 1-34 (cf. p. 56).

[Web+18] Raphaël Weber, Jingting Li, Catherine Soladié et Renaud Séguier. "A survey on databases of facial macro-expression and micro-expression". In : *International Joint Conference on Computer Vision, Imaging and Computer Graphics*. Springer. 2018, p. 298-325 (cf. p. 28).

[WRB11] Daniel Weinland, Remi Ronfard et Edmond Boyer. "A survey of vision-based methods for action representation, segmentation and recognition". In : *Computer vision and image understanding* 115.2 (2011), p. 224-241 (cf. p. 30).

[WKW16] Karl Weiss, Taghi M Khoshgoftaar et DingDing Wang. "A survey of transfer learning". In : *Journal of Big data* 3.1 (2016), p. 1-40 (cf. p. 27).

[Wen+21] Yu-Hui Wen, Zhipeng Yang et al. "Autoregressive stylized motion synthesis with generative flow". In : *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2021, p. 13612-13621 (cf. p. 48).

[Whe+15] Nkenge Wheatland, Yingying Wang et al. "State of the art in hand and finger modeling and animation". In : *Computer Graphics Forum*. T. 34. 2. Wiley Online Library. 2015, p. 735-760 (cf. p. 40).

[WR13] Sherri C Widen et James A Russell. "Children's recognition of disgust in others." In : *Psychological Bulletin* 139.2 (2013), p. 271 (cf. p. 58).

[Wig96] Jerry S Wiggins. *The five-factor model of personality : Theoretical perspectives*. Guilford Press, 1996 (cf. p. 21).

[Win83] David A Winter. "Biomechanical motor patterns in normal walking". In : *Journal of motor behavior* 15.4 (1983), p. 302-330 (cf. p. 95).

[WP95] Andrew Witkin et Zoran Popovic. "Motion warping". In : *Proceedings of the 22nd annual conference on Computer graphics and interactive techniques*. 00000. ACM, 1995, p. 105-108. URL : http://dl.acm.org/citation.cfm?id=218422 (visité le 19 nov. 2015) (cf. p. 47).

[Won+17] Jungdam Won, Jongho Park, Kwanyu Kim et Jehee Lee. "How to train your dragon : example-guided control of flapping flight". In : *ACM Transactions on Graphics (TOG)* 36.6 (2017), p. 1-13 (cf. p. 71).

[Wu+21] Jinting Wu, Yujia Zhang, Shiying Sun, Qianzhong Li et Xiaoguang Zhao. "Generalized zero-shot emotion recognition from body gestures". In : *Applied Intelligence* 52.8 (nov. 2021), p. 8616-8634. URL : https://doi.org/10.1007/s10489-021-02927-w (cf. p. 39).

[WTR11] Xiaomao Wu, Maxime Tournier et Lionel Reveret. "Natural Character Posing from a Large Motion Database". In : *IEEE Computer Graphics and Applications* 31.3 (mai 2011), p. 69-77 (cf. p. 102, 103).

[Wu+20] Zonghan Wu, Shirui Pan et al. "A comprehensive survey on graph neural networks". In : *IEEE transactions on neural networks and learning systems* 32.1 (2020), p. 4-24 (cf. p. 38).

[Xia+15a] Shihong Xia, Congyi Wang, Jinxiang Chai et Jessica Hodgins. "Realtime style transfer for unlabeled heterogeneous human motion". In : *ACM Transactions on Graphics (TOG)* 34.4 (2015). 00000, p. 119. URL : http://dl.acm.org/citation.cfm?id=2766999 (visité le 21 oct. 2015) (cf. p. 38, 46, 75).

[Xia+15b] Shihong Xia, Congyi Wang, Jinxiang Chai et Jessica Hodgins. "Realtime style transfer for unlabeled heterogeneous human motion". In : *ACM Transactions on Graphics (TOG)* 34.4 (2015), p. 1-10 (cf. p. 103).

[Xia+22] Weihao Xia, Yulun Zhang et al. "Gan inversion : A survey". In : *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2022) (cf. p. 48).

[YLv07] KangKang Yin, Kevin Loken et Michiel van de Panne. *SIMBICON : Simple Biped Locomotion Control*. Juill. 2007 (cf. p. 41).

[YM16] M. Ersin Yumer et Niloy J. Mitra. "Spectral style transfer for human motion between independent actions". en. In : *ACM Transactions on Graphics* 35.4 (2016). 00000, p. 1-8. url : http://dl.acm.org/citation.cfm?doid=2897824.2925955 (visité le 17 oct. 2016) (cf. p. 47, 67, 72, 103).

[Zac+13a] Haris Zacharatos, Christos Gatzoulis, Yiorgos Chrysanthou et Andreas Aristidou. "Emotion Recognition for Exergames Using Laban Movement Analysis". In : *Proceedings of Motion on Games*. MIG '13. Dublin 2, Ireland : Association for Computing Machinery, nov. 2013, p. 61-66 (cf. p. 38, 79).

[Zac+13b] Haris Zacharatos, Christos Gatzoulis, Yiorgos Chrysanthou et Andreas Aristidou. "Emotion recognition for exergames using laban movement analysis". In : *Proceedings of Motion on Games*. 2013, p. 61-66 (cf. p. 37).

[ZGC14] Haris Zacharatos, Christos Gatzoulis et Yiorgos L Chrysanthou. "Automatic emotion recognition based on body movement analysis : a survey". In : *IEEE computer graphics and applications* 34.6 (2014), p. 35-45 (cf. p. 35).

[Zad16] Kristjan Zadziuk. *Motion Matching, The Future of Games Animation... Today*. 2016 (cf. p. 45).

[Zaf+17] Stefanos Zafeiriou, Dimitrios Kollias et al. "Aff-wild : valence and arousal'In-the-Wild'challenge". In : *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*. 2017, p. 34-41 (cf. p. 29).

[Zen+18] Nianyin Zeng, Hong Zhang et al. "Facial expression recognition via learning deep sparse autoencoders". In : *Neurocomputing* 273 (2018), p. 643-649 (cf. p. 25).

[Zha+18a] Feifei Zhang, Tianzhu Zhang, Qirong Mao et Changsheng Xu. "Joint pose and expression modeling for facial expression recognition". In : *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2018, p. 3359-3368 (cf. p. 28).

[Zha+18b] He Zhang, Sebastian Starke, Taku Komura et Jun Saito. "Mode-adaptive neural networks for quadruped motion control". In : *ACM Transactions on Graphics (TOG)* 37.4 (2018), p. 1-11 (cf. p. 45).

[Zha+17] Kaihao Zhang, Yongzhen Huang, Yong Du et Liang Wang. "Facial expression recognition based on deep evolutional spatial-temporal networks". In : *IEEE Transactions on Image Processing* 26.9 (2017), p. 4193-4203 (cf. p. 27).

[ZT11] L. Zhang et D. Tjondronegoro. "Facial Expression Recognition Using Facial Movement Features". In : *IEEE Transactions on Affective Computing* 2.4 (2011), p. 219-229 (cf. p. 25).

[ZTC14] Ligang Zhang, Dian Tjondronegoro et Vinod Chandran. "Random Gabor based templates for facial expression recognition in images with facial occlusion". In : *Neurocomputing* 145 (2014), p. 451-464. URL : http://www.sciencedirect.com/science/article/pii/S0925231214005712 (cf. p. 22).

[Zha+22] Longwen Zhang, Chuxiao Zeng et al. "Video-driven Neural Physically-based Facial Asset for Production". In : *ACM Transactions on Graphics (TOG)* 41.6 (2022), p. 1-16 (cf. p. 91).

[ZP09] Guoying Zhao et Matti Pietikäinen. "Boosted multi-resolution spatiotemporal descriptors for facial expression recognition". In : *Pattern Recognition Letters* (2009). Image/video-based Pattern Analysis and HCI Applications. URL : http://www.sciencedirect.com/science/article/pii/S0167865509000695 (cf. p. 22, 52).

[Zha+21] Xibin Zhao, Junjie Zhu, Bingjun Luo et Yue Gao. "Survey on facial expression recognition : History, applications, and challenges". In : *IEEE MultiMedia* 28.4 (2021), p. 38-44 (cf. p. 30).

[Zha06] Li Zhaoping. "Theoretical understanding of the early visual processes by data compression and data selection". In : *Network : computation in neural systems* 17.4 (2006), p. 301-334 (cf. p. 52).

[Zin19] Fabio Zinno. *From Motion Matching to Motion Synthesis, and All the Hurdles In Between*. 2019 (cf. p. 45).