# Analyse des trajectoires sur une variété Grassmannienne pour la détection d'émotions dans les vidéos de profondeur

Taleb Alashkar[1] Boulbaba Ben Amor[1] Stefano Berretti[2] Mohamed Daoudi[1]

[1] Institut Mines-Télécom/Télécom Lille ; Laboratoire CRIStAL (UMR CNRS 9189), France.
[2] Department of Information Engineering, University of Florence, Italie.

Télécom Lille, rue G. Marconi, 59650 Villeneuve-d'Ascq, France.
taleb.alashkar@telecom-lille.fr

## Résumé

*Dans cet article nous présentons une nouvelle approche pour la détection des émotions humaines à partir de flux d'images de profondeur. Notre analyse consiste à découper la vidéo en sous-séquences chacune représentée par un sous-espace linéaire, élément d'une variété Grassmannienne. Il en résulte une trajectoire (courbe) sur la variété qui représente la vidéo à analyser. Les outils géométriques définis sur ce type de variété permettent de calculer une signature de l'évolution dans le temps de la dynamique (mouvements) de la personne filmée par la caméra. Cette signature est présentée au fur-et-à-mesure à un détecteur précoce d'événements, appelé SOSVM (Structured Output SVM), pour une analyse séquentielle. Les résultats obtenus sur la base publique Cam3D montrent l'intérêt de l'approche proposée pour l'analyse d'émotions spontanées filmées avec une caméra de bas-coût (type Kinect). Nos résultats montrent aussi que l'analyse des mouvements du corps est plus pertinente que l'analyse du visage seul dans le contexte des vidéos de profondeur.*

## Mots Clef

Émotions humaines, MS Kinect, trajectoires, variété Grassmannienne, SOSVM, image profondeur

## Abstract

*This paper proposes a new framework for online detection of spontaneous emotions from low-resolution depth sequences of the upper part of the body. To face the challenges of this scenario, depth videos are decomposed into subsequences, each modeled as a linear subspace, which in turn is represented as a point on a Grassmann manifold. Modeling the temporal evolution of distances between subsequences of the underlying manifold as a one-dimensional signature, termed Geometric Motion History, permits us to encompass the temporal signature into an early detection framework using Structured Output SVM, thus enabling online emotion detection. Results obtained on the publicly available Cam3D Kinect database validate the proposed solution, also demonstrating that the upper body, instead of the face alone, can improve the performance of emotion detection.*

## Keywords

Human emotions, MS Kinect, trajectories, Grassmann manifold, SOSVM. Depth Image

# 1 Introduction

Le déploiement de dispositifs dotés de caméras embarquées (e.g. les mobiles/tablettes, consoles de jeux, les ordinateurs personnels, les systèmes de vidéo-surveillance, etc.), a suscité l'intérêt des chercheurs en vision par ordinateur à développer des solutions de détection et de reconnaissance d'émotions des personnes filmées. Ces solutions trouvent leurs applications potentielles dans l'Interaction Homme-Machine (IHM) ou encore Homme/Robot, les jeux vidéo, la réalité augmentée et virtuelle, l'automobile (détection de la fatigue des conducteurs), etc. Les premières études sur les expressions faciales ont porté sur l'analyse d'images fixes [1], néanmoins, une expression faciale est une évolution dans le temps de l'état d'une personne ce qui implique une analyse des séquences vidéo, plutôt qu'une analyse dune image fixe. Par ailleurs, avec le développement de capteurs 3D, des chercheurs se sont tournés vers l'analyse de la forme 3D. Des approches récentes [2, 3] tentent de modéliser les expressions comme des déformations spatio-temporelles de la forme faciale 3D. Dans ce cas, les expressions faciales peuvent être étudiées en détail en analysant la dynamique temporelle de la forme faciale 3D (3D+t est souvent considéré comme données 4D ou des données 3D dynamique). Dans cette perspective, en plus de la richesse que peut présenter la forme pour analyser les déformations, son immunité relative aux variations d'illumination et la pose (orientation du visage) le rend encore plus plausible comparée à la 2D. Cet intérêt est d'autant plus important avec la démocratisation de caméras de profondeur bon marché, type Kinect, avec une résolution temporelle raisonnable (30 fps). Ce qui ouvre une voie prometteuse à de nouvelles opportunités et défis pour l'analyse des émotions humaines.

Le travail des psychologues, qui décrit les émotions humaines en matière d'espace discret (de catégories), a largement influencé la façon que la plupart des approches adoptent pour classer les émotions. L'exemple le plus répandu de cette catégorisation est donné par les six expressions prototypes (ou de base) de l'ensemble : *colère*, *dégoût*, *peur*, *joie*, *tristesse* et *surprise*. Cette description grossière a été spécialement renforcée par les études interculturelles menées par le psychologue pionnier Paul Ekman [4], indiquant que les humains perçoivent certaines émotions de base (par rapport aux expressions faciales) de la même manière, indépendamment de leurs cultures. L'influence de cette catégorisation des émotions (dites de base) est très visible dans les travaux menés jusqu'à présent en reconnaissance automatique d'émotions humaines. L'avantage d'une telle représentation est que les gens l'utilisent pour décrire les états émotionnels observés sur leurs pairs. Cependant, cette catégorisation discrète des émotions ne parvient pas à décrire un ensemble plus large d'émotions qui se produisent dans la communication naturelle en face-à-face. Bien que ces émotions de base soient les principaux points de référence de l'émotion, ils couvrent une partie d'un ensemble plus large de nos émotions quotidiennes.

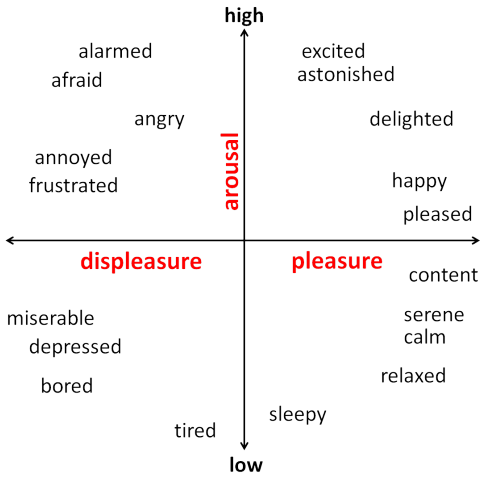


FIGURE 1 – Carte *Arousal-Valence* pour la description des émotions dans un espace continu. Des exemples d'émotions sont placées sur cet espace en fonction de l'état (positif/négatif) et de l'activation (passif/actif).

Une alternative à la *description catégorique* des émotions humaines est la *description dimensionnelle* [5], dans laquel un état émotionnel s'exprime en termes d'un petit nombre de dimensions, plutôt que dans un ensemble de catégories d'émotion. En particulier, les modèles les plus utilisés comportent les dimensions *Arousal-Valence* qui sont censés refléter les principaux aspects de l'émotion. La dimension *Valence* mesure comment un être humain se sent en adoptant une échelle allant du positif au négatif, en revanche, la dimension *Arousal* mesure si les humains sont plus ou moins susceptibles de prendre une action en vertu de l'état émotionnel, allant de l'actif au passif [6]. Une description dimensionnelle des émotions est affichée sur la figure 1, en utilisant la carte *Arousal-Valence*. Sur l'axe horizontal, la dimension Valence est quantifiée du *displeasure* au *pleasure*, alors que sur l'axe vertical, la dimension *Arousal* est représenté par l'état d'éveil, variant du niveau faible *low* au niveau *high*. Dans les systèmes automatiques de reconnaissance des émotions qui sont basés sur la représentation de l'émotion dans l'espace *Arousal-Valence*, le problème est souvent encore simplifiée à une classification binaire en considérant les classes (positif/négatif et actif/passif) ou encore quatre classes (quadrants de l'espace présenté par la Fig. 1).

En plus des limitations posées par la classification catégorielle, la plupart des solutions actuelles pour la reconnaissance de l'expression du visage à partir de séquences dynamiques 3D sont évaluées dans des scénarios contraints [7], encore loin des données du monde réel. Il s'agit de séquences de visages 3D de haute résolution d'expressions posées (non-spontanées), par exemple la base BU-4DFE [2]. La reconnaissance des expressions faciales spontanées est un problème plus difficile qui a récemment attiré l'intérêt de la communauté de la vision par ordinateur (voir par exemple les travaux publiés dans [8] et [9]). L'effet de la faible résolution d'acquisitions bruitées sur la reconnaissance d'expression n'a pas été pris en compte dans ces études. La majorité des méthodes proposées pour la classification des expressions sont basées sur l'observation des séquences dans leur globalité (i. e., la décision est prise une fois la séquence complète observée). Très peu d'intérêt a été porté au temps minimum nécessaire pour prendre une décision, ce qu'on appelle aussi classification précoce. La classification précoce est d'une grande importance dans plusieurs applications, tels que les jeux vidéo, la détection de chutes, etc. Le compromis entre la taille de l'observation et la précision du détecteur/classificateur est le point à étudier à travers une analyse séquentielle des vidéos. Dans ce contexte, très récemment, Hoai et De la Torre [13] ont proposé une formulation d'apprentissage pour la détection précoce d'événements. Leur cadre appelé "max-margin early event detector" est conçu pour l'apprentissage de détecteurs d'événements temporels capables de reconnaître des événements partiels, permettant ainsi la détection précoce avec une latence minimale. Leur méthode étend l'algorithme SOSVM (Structured Output Support Vector Machine) aux données séquentielles. Ils ont illustré ce cadre sur la détection des expressions faciales dans des images 2D, la reconnaissance des gestes de la main et la classification des activités humaines à partir de séquences vidéo.

Dans la littérature de la reconnaissance des émotions, un autre aspect rarement pris en compte est la pertinence de la dynamique du corps, en plus des expressions faciales, pour transmettre des émotions [14]. En particulier, quelques études de différents domaines ont convenu que la combinaison des expressions du visage et du corps peut améliorer

la reconnaissance des états émotionnels [15, 16]. La prise en compte conjointe de ces aspects est désormais encouragée par la promotion des technologies de l'acquisition, qui permet la capture de données 3D de la profondeur du corps ainsi que le visage. Cette approche a récemment été utilisée pour améliorer la compréhension de l'Interaction Homme-Machine [17].

Partant des constats cités ci-dessus, cet article propose une approche de détection précoce, capable de reconnaître les émotions avec une faible latence. La solution proposée est appliquée à un scénario difficile, où les séquences de profondeur de la partie supérieure du corps sont acquises avec un capteur à faible résolution (du type Kinect). Par ailleurs, les émotions capturées sont naturelles et spontanées ce qui implique à la fois des expressions du visage et des mouvements corporels. Le reste du papier est organisé comme suit. La Section 2 présente les principales idées et les contributions de l'approche proposée. Dans la Section 3, la représentation proposée d'une séquence vidéo comme un ensemble de points sur une variété Grassmannienne est présentée. Nous rappelons que : (1) une Grassmannienne $\mathcal{G}_k$ de $\mathbb{R}^n$ est l'ensemble des sous-espaces de $\mathbb{R}^n$ de dimension $k$ ; (2) une variété de Stiefel est l'ensemble des matrices de taille $n \times k$ dont les colonnes sont orthogonales et unitaires (pour plus de détails sur ces variétés nous renvoyons le lecteur à [23]). Dans la Section 4, la représentation 3D de la séquence dynamique est adaptée à un cadre de détection précoce d'événement, inspiré des travaux de Hoai et De la Torre [13]. Le potentiel de la solution proposée est présenté dans la Section 5, en illustrant les résultats obtenus sur la base de données Cam3D Kinect. Enfin, quelques conclusions et perspectives sont discutées dans la Section 6.

## 2 Vue d'ensemble et contributions

Dans cet article, nous proposons une approche de détection d'émotions à partir de l'analyse de flux d'images de profondeur de la partie supérieure du corps filmée par une Kinect. À cette fin, nous couplons deux approches complémentaires du domaine de la reconnaissance des formes – (1) une analyse dynamique des données en utilisant des outils de géométrie différentielle ; et (2) la détection précoce d'événement en utilisant une adaptation de l'algorithme SOSVM pour l'analyse séquentielle de ces données. Une vidéo de profondeur (séquence temporelle de trames de profondeur) est d'abord divisée en sous-séquences d'un nombre prédéfini de trames adjacentes. Chaque sous-séquence est transformée en une représentation matricielle pour présenter un sous-espace linéaire (c-à-d une matrice orthonormée dont les vecteurs colonnes engendrent un sous-espace linéaire). Un sous-espace (et par conséquent une sous-séquence) est tout naturellement vu comme un élément d'une variété Grassmannienne. Ensuite, une des métriques appropriées définie sur cette variété est employée pour évaluer les divergences entre des éléments représentants différentes sous-séquences. Dans la Figure 2, nous illustrons l'idée de faire correspondre les sous-séquences d'une vidéo de profondeur à une variété Grassmannienne. Ainsi, au fur et à mesure de l'acquisition un nouveau point vient s'ajouter aux éléments précédemment obtenus, ce qui peut être vu comme une trajectoire (une courbe) sur la variété. Nous examinons ensuite l'évolution temporelle des distances entre les points voisins sur cette trajectoire, ce qui donne un vecteur mono-dimensionnel qui a pour vocation de capturer la dynamique (les mouvements) du corps filmé par la caméra. L'analyse séquentielle consiste donc à capturer l'historique de la dynamique humaine en face de la caméra qui servira de description pour détecter l'émotion. Cette dernière étape consiste à présenter la description obtenue à l'instant $t$ à un cadre de la détection précoce d'événements permettant ainsi la détection en ligne et par conséquent la définition d'un temps d'arrêt pour l'analyse.

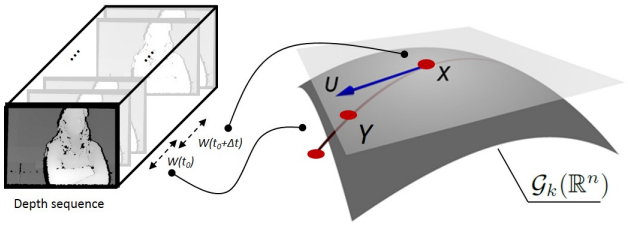


FIGURE 2 – Représentation d'une vidéo de profondeur par une trajectoire sur la variété Grassmannienne $\mathcal{G}_k(\mathbb{R}^n)$.

En résumé, les principales contributions de ce travail sont :

- Une méthode pour représenter une acquisition de profondeur sous forme d'une trajectoire sur une variété Grassmannienne ;
- Une méthode d'extraction d'une signature temporelle permettant de capturer la dynamique de la partie supérieure du corps (y compris de la tête/visage) ;
- Une analyse séquentielle de la signature temporelle extraite par une adaptation du classificateur SOSVM, pour une détection précoce des émotions humaine.

Nous soulignons également que le cadre proposé est le premier, à notre connaissance, capable de détecter des émotions spontanées dans un scénario complexe qui comprend :

- Des flux d'images de profondeur de la partie supérieure du corps acquise avec une caméra de profondeur grand public, la Kinect ;
- Des émotions spontanées acquises sans un protocole rigide (c.-à-d., aucune hypothèse sur le moment où l'émotion se produit dans la séquence) ;
- Les émotions liées non seulement à la dynamique temporelle des déformations du visage 3D, mais aussi à la posture et au mouvement de la tête et de la partie supérieure du corps.

Dans la section qui suit, nous motivons nos choix d'analyser des images de profondeur et de la représentation de ces

vidéos sur des variétés de Stiefel et de Grassmann.

# 3 Trajectoires sur les variétés de Stiefel et de Grassmann

Un des avantages derrière l'utilisation de flux d'images de profondeur est le fait qu'il est relativement facile d'isoler et de suivre le corps humain dans la scène observée. En outre, les cartes de profondeur sont indépendantes des changements d'illumination et fournissent une représentation plus complète de la forme du corps humain et de sa dynamique. Malgré les limitations dues aux données bruitées et de faible résolution spatiale, l'analyse de ces données temporelles dans des cadres d'analyse adaptés à la nature de ces données, peut être une solution de l'analyse temporelle car le corps est une surface 3D déformable au fil du temps. Ainsi, l'introduction de la dimension temporelle peut être utile pour améliorer la reconnaissance statique. Rappelons que cette tendance à analyser des données dynamiques (vidéos) est maintenant bien établie dans le domaine 2D. Pour surmonter les limitations mentionnées ci-dessus (de bruits dans les données de faibles résolution spatiale), notre idée est d'adopter une représentation sous forme de sous-espaces linéaires (souvent représentés par des matrices orthonormées) et profiter des outils de la géométrie pour analyser ces données. D'abord, le fait de considérer une sous-séquence de quelques trames successives de la vidéo de profondeur peut être vu comme une phase d'amélioration de la résolution spatiale. Le second avantage est de réduire l'effet du bruit des acquisitions en ne gardant que les larges vecteurs propres de la matrice obtenue par SVD (Décomposition en Valeurs Singulières).

Plus formellement, après avoir isolé le corps humain de l'arrière-plan dans les images de profondeur et normalisé le nombre de pixels à $n$ sur cette partie, comme une étape de pré-traitement, une fenêtre de trames successives $W(t_0)$ ($t_0$ correspond au premier instant de l'acquisition) est mise en correspondance avec une variété de *Stiefel* [19]. Nous appliquons sur $W(t_0)$, une fois transformée en représentation matricielle où chaque colonne représente une image, la méthode $k$-SVD pour produire une matrice orthonormale. Pour une représentation dans la variété de Grassmann, la sous-séquence est modélisée par un sous-espace engendré par les larges vecteurs propres de la matrice orthonormale produite précédemment. La même procédure est appliquée aux fenêtres de trames $W(t_0 + i\Delta t)$ vu à l'instant $t_0 + i\Delta t$, où $i \in \{1, \ldots, T\}$. En conséquence, la vidéo de profondeur est mise en correspondance avec la variété de Stiefel ($\mathcal{V}_k(\mathbb{R}^n)$) et peut-être vue comme une une trajectoire (voir figure 2.), elle est aussi mappée sur la variété de Grassmann $\mathcal{G}_k(\mathbb{R}^n)$ en une trajectoire. Le problème de la représentation dans $\mathcal{V}_k(\mathbb{R}^n)$ est que les vecteurs colonnes de deux matrices données $M$ et $M'$ peuvent engendrer le même sous-espace. Contrairement à la variété de Stiefel, des points sur la variété Grassmannienne $\mathcal{G}_k(\mathbb{R}^n)$ sont des classes d'équivalence de points dans $\mathcal{V}_k(\mathbb{R}^n)$, où deux éléments sont équivalents si leurs vecteurs colonnes engendrent le même sous-espace. En d'autres termes, on peut écrire $\mathcal{G}_k(\mathbb{R}^n)$ est un espace quotient de $\mathcal{V}_k$ ($\mathcal{G}_k(\mathbb{R}^n) = \mathcal{V}_k(\mathbb{R}^n)/O(k)$, où $O(k)$ est le groupe orthogonal de dimension $k$. Autrement, $\mathcal{G}_k(\mathbb{R}^n)$ est l'ensemble des orbites de $\mathcal{V}_k(\mathbb{R}^n)$ sous l'action du groupe $O(k)$.

Afin de quantifier la distance entre les points sur la variété de Stiefel ou de Grassmann il est nécessaire de définir des métriques appropriées. Prenons $Y_1, Y_2 \in \mathcal{V}_k(\mathbb{R}^n)$ et $\mathcal{Y}_1 = Span(Y_1)$, $\mathcal{Y}_2 = Span(Y_2) \in \mathcal{G}_k(\mathbb{R}^n)$. Les métriques peuvent être définies comme suit,

- **Métrique sur la variété de Stiefel** : La métrique de Frobenius définie par $d_{\mathcal{V}}(Y_1, Y_2) = \|Y_1 - Y_2\|_F$, où $\| \ \|_F$ est la norme matricielle de Frobenius.
- **Métrique sur la variété de Grassmann** : Golub *et al.* [18] ont introduit de manière intuitive et efficace une distance entre deux sous-espaces linéaires en utilisant les angles principaux. L'ensemble d'angles principaux $\Theta = [\theta_1, \ldots, \theta_k]$ entre $\mathcal{Y}_1$ et $\mathcal{Y}_2$ est défini comme suit :

$$\theta_i = \cos^{-1}\left(\max_{u_i \in \mathcal{Y}_1} \max_{v_i \in \mathcal{Y}_2} \langle u_i^t, v_i \rangle\right), \qquad (1)$$

où $u$ et $v$ sont les vecteurs colonnes qui engendrent les espaces linéaires, $\mathcal{Y}_1$ et $\mathcal{Y}_2$ respectivement, sous réserve des contraintes supplémentaires $\langle u^t, u \rangle = \langle v^t, v \rangle = 1$, et $\langle u^t, v \rangle = \langle v^t, u \rangle = 0$, où $\langle , \rangle$ désigne le produit scalaire dans $\mathbb{R}^n$. Basé sur la définition des angles principaux 1, la distance géodésique entre $\mathcal{Y}_1$ et $\mathcal{Y}_2$ peut être définie par $d_{\mathcal{G}}(\mathcal{Y}_1, \mathcal{Y}_2) = \sqrt{\sum_i \theta_i^2}$ [19].

Les métriques $d_{\mathcal{V}}$ et $d_{\mathcal{G}}$ sont utilisées pour calculer la *Geometric Motion History* en analysant de manière séquentielle les trajectoires sur les variétés de Stiefel et de Grassmann, respectivement. Nous notons que la représentation de données sur ces variétés ont été déjà utilisées avec succès dans la littérature et appliquée à la reconnaissance 2D de l'activité humaine [20], l'estimation de l'âge [21], et la reconnaissance de visages [22] à partir de séquences vidéo ou d'ensembles d'images fixes (pour un survol plus complet, nous renvoyons le lecteur à [24]). Malgré la base mathématique commune, notre méthodologie est très différente par rapport aux études mentionnées ci-dessus, car il s'agit de mapper les vidéos de profondeur pour former des trajectoires sur les variétés sur lesquelles on prévoit une analyse séquentielle. Ceci permet d'avoir une décision avec un minimum de latence (par rapport aux observations temporelles complètes [20]), et définissant ainsi un temps d'arrêt.

# 4 Détection précoce d'émotions

La tâche de détection d'une émotion dans une vidéo de profondeur est formulée ici comme un problème de détection précoce, qui vise à détecter l'émotion d'intérêt aussi vite que possible. A cet effet, nous utilisons le classificateur SOSVM (*Structured Output Support Vector Ma-*

*chine*), qui se présente comme un problème d'optimisation convexe [25]. Les principales motivations d'utilisation du SOSVM sont les suivants : (1) il peut être entraîné sur tous les segments partiels et complets d'une vidéo à la fois ; (2) il permet de modéliser la corrélation entre les caractéristiques extraites et la durée de l'émotion ; (3) aucune connaissance préalable n'est requise sur la structure de l'émotion ; (4) il a donné de meilleures performances que d'autres détecteurs [26].

### 4.1 Extraction des *Geometric Motion History* : GMH

Les représentations des vidéos de profondeurs par des trajectoires, de matrices orthonormales ou encore de sous-espaces linéaires, sur des variétés de Stiefel ou de Grassmann, nous permettent d'utiliser les outils géométriques pour calculer les distances entre les points et ainsi quantifier la différence entre les sous-séquences successives. Afin d'extraire les caractéristiques temporelles de la séquence à analyser (par conséquent capturer la dynamique), l'idée est de calculer séquentiellement les distances entre les points successifs et de construire un historique de la dynamique du corps humain en interaction. Plus formellement, étant donné une trajectoire $\mathcal{T}$ de sous-espaces de dimension $k$ (de matrices orthonormales $n \times k$) de $\mathbb{R}^n$, $\{\mathcal{X}_i\}_{i\in\{0,\ldots,T\}}$, nous calculons la longueur du chemin géodésique connectant $\mathcal{X}_{i+1}$ à $\mathcal{X}_i$, que nous ajoutons à l'historique déjà calculé en analysant la fraction vue de la vidéo. Ceci résulte en un signal à une dimension variant au cours du temps que nous appelons *Geometric Motion History*, comme illustré sur la figure 3. En particulier, dans cette figure les graphes montrent les vecteurs caractéristiques *GMH* pour trois segments vidéos concaténés, où le segment vert correspond à l'émotion d'intérêt (*joie* ici), qui est comprise entre deux autres segments vidéos de deux émotions qui représentent des émotions autres que la joie.

### 4.2 Analyse séquentielle par SOSVM *(Structured Output SVM)*

Prenons un ensemble de vecteurs caractéristiques *GMH* notés par, $v_1, \ldots, v_n$ de différentes émotions que nous concaténons, comme le montre la figure 3. Chaque vecteur caractéristique, $v_i$, comprend une seule émotion, qui est annoté par deux valeurs $(s^i, e^i)$, pour définir le début et la fin de l'émotion, respectivement. A tout instant $t^i$ de la vidéo de l'émotion ($s^i \leq t^i \leq e^i$), toutes les *GMH* partielles obtenus entre $[0, t^i]$ seront utilisés pour entraîner le classificateur SOSVM. Les $v_i$ de l'émotion d'intérêt (à détecter) sont étiquetés par $+1$ alors que les autres $v_i$ sont étiquetés par $-1$. Dans la phase de test, on attend du SOSVM qu'il décide s'il s'agit de l'émotion d'intérêt ou non, et ceci dès que possible (après son démarrage et avant que l'émotion ne se termine). Nous adoptons une méthodologie pour l'apprentissage des données séquentielles par SOSVM inspirée des travaux présentés dans [13].

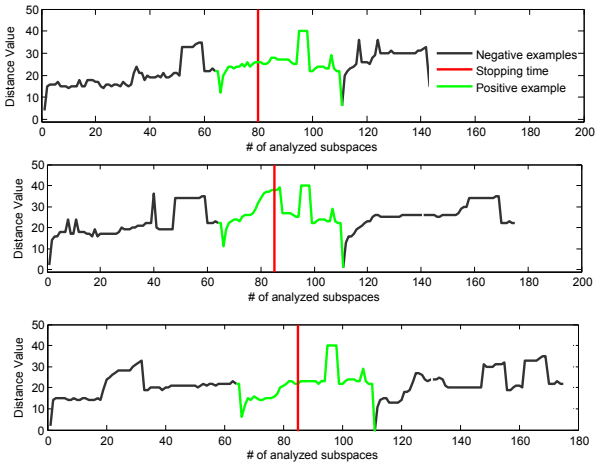


FIGURE 3 – Trois exemples de la *Geometric Motion History*, signature temporelle extraite par le cadre proposé. On peut observer en vert l'émotion d'intérêt ici (la *Joie* entourée de part et d'autre par une signature d'une autre émotion autre que la *Joie*.En rouge, on représente le moment auquel la décision sur l'émotion d'intérêt est disponible.

## 5 Expérimentations

L'approche proposée a été évaluée sur la base de données Cam3D Kinect [27] en utilisant différents scénarios et différents paramètres. Dans cette base de données, *Mahmoud et al* [27] ont recueilli un ensemble de 108 segments audio/vidéo d'états mentaux complexes naturels de 7 sujets. Chaque vidéo est acquise avec une Kinect, incluant à la fois l'apparence (color) et la profondeur (depth). Les états émotionnels annotés sont : *Agreeing*, *Bored*, *Disagreeing*, *Disgusted*, *Excite*, *Happy*, *Interested*, *Sad*, *Surprised*, *Thinking* and *Unsure*. Ces états émotionnels sont plus réalistes et plus complexes que les émotions de base, bien connues dans la littérature. Le tableau 1 indique le nombre de segments disponibles pour chaque état émotionnel ou état mental complexe.

TABLE 1 – Nombre de vidéos disponibles dans la base Cam3D pour chaque état mental complexe [27].

| **Émotion/État mental** | **# de segments** |
|---|---|
| Accord (Agreeing) | 4 |
| Ennuyé (Bored) | 3 |
| Désaccord (Disagreeing) | 2 |
| Dégoûté (Disgusted) | 1 |
| Excité (Excited) | 1 |
| **Joie (Happy)** | **26** |
| Intérêt (Interested) | 7 |
| Neutre (Neutral) | 2 |
| Triste (Sad) | 1 |
| Surprise (Surprised) | 5 |
| **Réflexion (Thinking)** | **22** |
| **Incertitude (Unsure)** | **32** |

On peut remarquer que les vidéos dans cet ensemble de données fournissent un échantillon d'émotions décrites par la carte d'émotions de l'espace *Arousal-Valence* comme indiqué dans la figure 1. Cependant, la possibilité d'utiliser chaque catégorie d'émotion dans une expérience de détection est entravée par le faible nombre de vidéos dans certaines classes (9 des 12 classes ne comportant que au plus 8 vidéos). Ceci nous a motivé à examiner les deux scénarios expérimentaux suivants : la détection de la *joie* d'une part ; et la détection de *réflexion (Thinking)/Incertitude (Unsure)* d'un autre côté. Par rapport à la carte de la figure 1, le premier scénario teste la détection d'une émotion située dans le *haute excitation/plaisir* quadrant (d'émotion positive) ; le second fait référence à une émotion dans le secteur *faible excitation/déplaisir* (émotion négative).

Deux critères d'évaluation différents sont utilisés pour tester les performances du point de vue de la précision et de la rapidité et de la détection de l'émotion de l'émotion d'intérêt : (1) **Aire sous la courbe ROC :** Une courbe ROC est créée en traçant le taux de vrais positifs (TPR) par rapport au taux de faux positifs (FPR) en faisant varier le seuil ; (2) **Courbe AMOC :** La courbe AMOC (Activity Monitoring Operating Characteristic) représente un score de rapidité (Normalized Time to Detect) en fonction du taux de fausses alertes (FPR), pour différents seuils de détection. Nous avons appliqué l'approche proposée pour détecter les états émotionnels de deux régions différentes de l'espace des émotions de la figure 1 : (1) *Joie* (figurant sur le quadrant haute excitation/plaisir) ; (2) *Réflexion/Incertitude* (figurant dans le cadrant faible excitation/déplaisir). L'émotion d'intérêt et d'autres segments dans les deux expériences sont répartis de façon égale en apprentissage et en test. Puis, un vecteur caractéristique *GMH* d'un segment de l'émotion d'intérêt est concaténé à deux vecteurs caractéristiques *GMH* calculés sur d'autres émotions. La figure 3 illustre cette opération de concaténation de signatures temporelles, en vert l'émotion d'intérêt (donc à détecter), et en noir (de part et d'autre) les signatures des émotions choisies aléatoirement d'autres classes. L'objectif est donc de détecter l'émotion d'intérêt dans cette concaténation de vidéos étiquetées par leurs groupes d'appartenance. Nous formons un un total de 100 *GMH* pour l'apprentissage, et le même nombre pour le test. Pour chaque séquence générée, le début et la fin de l'émotion d'intérêt est connus. Les distances calculées sur les variétés de Stiefel ou de Grassmann (voir Sec 3) sont extraites pour comparaison. L'effet de la taille de fenêtre utilisée pour l'extraction des sous-séquences est également analysé.

Pour la détection de l'émotion *Joie*, la figure 4 montre les courbes ROC et les courbes de AMOC obtenues en analysant séquentiellement les *GMH*, signatures calculées pour les représentations dans les variétés de Grassmann (en rouge) et de Stiefel (en bleue). Notons que nous moyennons les résultats de 20 expériences différentes. A partir des courbes ROC relatives à la représentation dans la Grassmannienne, on peut voir que, lorsque le FPR est d'environ 20% le $TPR$ atteint 90% de détection de *Joie*. Cette précision diminue de manière significative (environ 50%) pour un FAR = 10%. En comparant l'analyse des trajectoires le long de la variété de Stiefel (Les courbes en tirets) à l'analyse des trajectoires sur la variété de Grassmann (courbes continues), il ressort clairement que la seconde surpasse la première. Les aires sous les courbes ROC sont 0.73 et 0.84 sur Stiefel et sur Grassmann, respectivement. Cela démontre la pertinence de l'utilisation des sous-espaces $\mathcal{Y} = Span(Y)$ et la métrique associée $d_{\mathcal{G}}$. Ceci est principalement dû à l'invariance de la représentation de sous-espaces aux rotations (éléments du groupe $O(k)$) et donc du fait que $\mathcal{G}$ est un espace de quotient de $\mathcal{V}$ sous l'action du groupe $O(k)$. Les courbes sur la droite de la figure 4 montrent l'évolution de la latence du système (la fraction de la vidéo nécessaire pour prendre la décision binaire) contre le FPR. Par exemple, le détecteur atteint 20% du FPR en analysant 20% du segment vidéo. Une fois de plus, les résultats obtenus avec les représentations sur la variété de Grassmann sont meilleurs comparés à ceux obtenus via la représentation sur la variété de Stiefel.

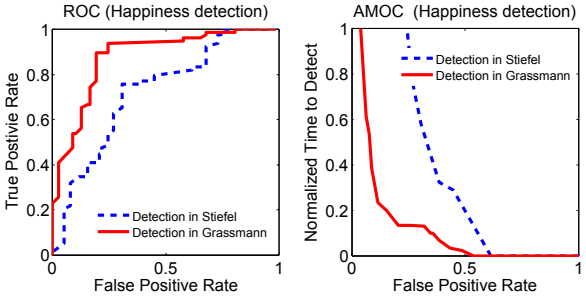


FIGURE 4 – Courbes ROC et AMOC pour la détection de *Happiness* sur les variété de Stiefel et de Grassmann.

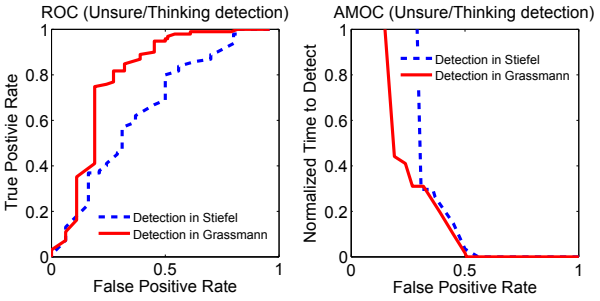


FIGURE 5 – Courbes ROC and AMOC pour la détection de *Thinking/Unsure* sur les variété de Stiefel et de Grassmann.

Dans une seconde expérience, la précision de la détection pour l'état affectif *Réflexion/Incertitude* est considérée. Les résultats rapportés par la figure 5, montrent une diminution

de la performance par rapport aux résultats de la détection de l'émotion *Joie*. Les aires sous les courbes ROC sont $0.66$ et $0.79$ sur les variétés de Stiefel et de Grassmann, respectivement. Ces résultats confirment encore une fois l'intérêt de la représentation par des trajectoires sur la variété de Grassmann. Les courbes AMOC présentées sur la droite de la figure 5 montrent que $20\%$ des échantillons négatifs sont reconnus comme étant des éléments de cette classe, même en analysant la totalité de l'observation (Normalized Time to Detect = 100%). Cela peut être expliqué par le "comportement" neutre commun présenté par les êtres humains lors d'états mentaux complexes de type par exemple, acceptation, ennuyé, reflexion, incertitude, etc. Cela induit le détecteur en erreur, ce qui n'était pas le cas pour le détecteur de joie précédemment étudié, car l'émotion de joie est souvent accompagnée de mouvements du corps et d'expressions faciales.

Pour approfondir sur l'importance d'utiliser la partie supérieure du corps plutôt que d'utiliser seulement le visage, nous avons aussi utilisé le même protocole précédent défini sur la base de données après le recadrage de la région de visage. De la figure 6, il est clair que la partie supérieure du corps est plus informative que le visage dans la transmission de l'émotion d'intérêt, dans un contexte d'acquisition par une Kinect. Dans l'expérience *Joie*, l'aire sous la courbe ROC pour la partie supérieure du corps et du visage sont $0.84$ et $0.68$, respectivement. On notera la même remarque, pour l'émotion d'intérêt *Réflexion/Incertitude*, où les aires sous les courbes ROC sont $0.79$ et $0.63$ pour la partie supérieure du corps et du visage, respectivement.

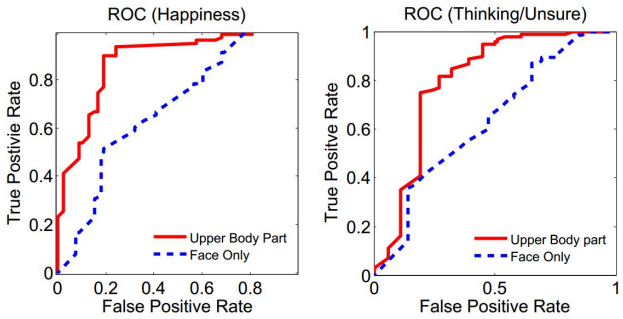


FIGURE 6 – Courbes ROC pour la détection des émotions *Joie* et *Réflexion/Incertitude* sur une variété Grassmannienne en se basant sur la dynamique du visage et du corps.

# 6 Conclusions

Dans cet article, nous avons présenté une nouvelle signature temporelle qui caractérise la dynamique du corps humain et son utilisation pour la détection des émotions. Notre idée est de considérer des espaces d'émotions continues bien établis (par exemple, *Arousal-Valence*) et de définir une région (émotion) d'intérêt pour la détection automatique. Les vidéos de profondeurs issues d'une caméra Kinect sont d'abord mises en correspondance avec une variété de Grassmann pour pallier la médiocre qualité des données (basse résolution, données manquantes, et le bruit). Il en résulte des trajectoires sur cette variété à analyser pour la détection d'une émotion d'intérêt. Un classificateur dédié à l'analyse séquentielle (SOSVM) nous permet d'étudier le compromis entre la précision de la détection et le temps de latence du système. Notre approche a été testée sur le nouveau jeu de données Cam3D Kinect, qui comprend un nombre limité de vidéos annotées et segmentées. Les résultats obtenus ont montré l'intérêt de l'approche proposée, illustrée sur deux états émotionnels spécifiques (Joie et Réflexions/Incertitude). Ils ont montré également un net avantage de l'utilisation de l'expression de la partie supérieure du corps par rapport à l'analyse du visage, dans le contexte de l'acquisition par des caméras de profondeur à bas-coût.

## Remerciements

Les auteurs tiennent à remercier les auteurs de [13] pour leur avoir fourni les codes de SOSVM ainsi que les outils d'évaluation. Ce travail a reçu le financement du projet MAGNUM 2 (BPI et Région Nord-Pas de Calais).

# Références

[1] M. Pantic and L. J. M. Rothkrantz, "Automatic analysis of facial expressions : The state of the art," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol.22, pp.1424–1445, 2000.

[2] Y. Sun, X. Chen, M. Rosato and L. Yin, "Tracking vertex flow and model adaptation for 3D spatio-temporal face analysis," *IEEE Trans. on Systems, Man, and Cybernetics – Part A*, vol.40, pp.461–474, 2010.

[3] B. Ben Amor, H. Drira, S. Berretti, M. Daoudi, and A. Srivastava, "4–D facial expression recognition by learning geometric deformations, *IEEE Trans. on Cybernetics*, vol.44, no.12, pp.2443,2457, Dec. 2014.

[4] P. Ekman, "Universals and cultural differences in facial expressions of emotion," in *Nebraska Symposium on Motivation*, Lincoln, vol.19, pp.207–283, 1972.

[5] J. Russell and A. Mehrabian, "Evidence for a three-factor theory of emotions," *Jour. of Research in Personality*, vol.11, pp.273–294, 1977.

[6] A. Vinciarelli, M. Pantic and H. Bourlard, "Social signal processing : Survey of an emerging domain," *Image and Vision Computing Jour.*, vol.27, pp.1743–1759, 2009.

[7] G. Sandbach, S. Zafeiriou, M. Pantic and L. Yin, "Static and dynamic 3D facial expression recognition : A comprehensive survey," *Image and Vision Computing*, vol.30, pp.683–697, 2012.

[8] S. Wan and J. Aggarwal, "Spontaneous facial expression recognition : A robust metric learning approach," *Pattern Recognition Jour.*, vol.47, pp.1859–1868, 2014.

[9] M. Abd El Meguid and M. Levine, “Fully automated recognition of spontaneous facial expressions in videos using random forest classifiers," *IEEE Trans. on Affective Computing*, vol.5, pp.141–154, 2014.

[10] K. Schindler and L. Van Gool, “Action snippets : How many frames does human action recognition require ?," in *IEEE Conf. on Computer Vision and Pattern Recognition*, pp.1−8, 2008.

[11] L. Su, S. Kumano, K. Otsuka, D. Mikami, J. Yamato and Y. Sato, “Early facial expression recognition with high-frame rate 3D sensing," in *IEEE Int. Conf. on Systems, Man, and Cybernetics*, pp.3304–3310, 2011.

[12] L. Su and Y. Sato, “Early facial expression recognition using early rankboost," in *IEEE Int. Conf. on Automatic Face and Gesture Recognition*, pp.1–7, 2013.

[13] M. Hoai and F. De la Torre, “Max-margin early event detectors," *Int. Jour. of Computer Vision*, vol.107, pp.191−202, 2014.

[14] A. Kleinsmith and N. Bianchi-Berthouze, “Affective body expression perception and recognition : A survey," *IEEE Trans. on Affective Computing*, vol.4, pp.15−33, 2013.

[15] J. Van den Stock, R. Righart and B. de Gelder, “Body expressions influence recognition of emotions in the face and voice," *Emotion Jour.*, vol.7, pp.487−494, 2007.

[16] H. Meeren, C. van Heijnsbergen and B. de Gelder, “Rapid perceptual integration of facial expression and emotional body language," in *National Academy of Sciences*, USA, vol.45, pp.16518−16523, 2005.

[17] N. Bianchi-Berthouze, “Understanding the role of body movement in player engagement," in *Human Computer Interaction*, vol.28, pp.40−75, 2013.

[18] G. H. Golub and C. F. Van Loan, “Matrix Computations,” (3rd Ed.), *Johns Hopkins University Press*, Baltimore, MD, USA, 1996.

[19] A. Edelman, T. A. Arias and S. T. Smith, “The geometry of algorithms with orthogonality constraints," *SIAM J. Matrix Anal. Appl*, vol.20, pp.303–353, 1998.

[20] P. K. Turaga, A. Veeraraghavan, A. Srivastava and R. Chellappa, “Statistical computations on grassmann and stiefel manifolds for image and video based recognition," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol.33, pp.2273–2286, 2011.

[21] T. Wu, P. K. Turaga and R. Chellappa, “Age estimation and face verification across aging using landmarks," in *IEEE Transactions on Information Forensics and Security*, vol.7, pp.1780−1788, 2012.

[22] J. Hamm and D. D. Lee, “Grassmann discriminant analysis : A unifying view on subspace-based learning," in *Int. Conf. on Machine Learning*, vol.08, pp.376−383, 2008.

[23] A. Edelman, R. Arias, S. Smith, The geometry of algorithms with orthogonal constraints, SIAM Journal on Matrix Analysis and Applications 2 (1999). 303–353

[24] Y. M. Lui, “Advances in matrix manifolds for computer vision," *Image Vision Computing Jour.*, vol.30, pp.380−388, 2012.

[25] I. Tsochantaridis, T. Joachims, T. Hofmann and Y. Altun, “Large margin methods for structured and interdependent output variables," *Jour. of Machine Learning Research*, vol.6, pp.1453−1484, 2005.

[26] N. Nguyen and Y. Guo, “Comparisons of sequence labeling algorithms and extensions," in *Int. Conf. on Machine Learning*, pp.681−688, 2007.

[27] M. Mahmoud, T. Baltrŭsaitis, P. Robinson and L. Riek, “3D corpus of spontaneous complex mental states," in *Conf. on Affective Computing and Intelligent Interaction*, pp.205−214, 2011.