Détection de la dépression par l'analyse de la géométrie faciale et de la parole

Anastasia PAMPOUCHIDOU[1], Olympia SIMANTIRAKI[2], Calliope-Marina VAZAKOPOULOU[3], Kostas MARIAS[2, 3], Panagiotis SIMOS[4], Fan YANG[1], Fabrice MERIAUDEAU[1, 5], Manolis TSIKNAKIS[2, 3]

[1]Le2i Laboratory, University of Burgundy, Le Creusot, France

[2]Foundation for Research & Technology - Hellas, Heraklion, Greece

[3]Technological Educational Institute of Crete, Department of Informatics Engineering, Heraklion, Greece

[4]Department of Psychiatry, University of Crete, Heraklion, Crete, Greece

[5]CISIR, Electrical Engineering Department, Universiti Teknologi PETRONAS, Malaysia

{anastasia.pampouchidou, fan.yang, fabrice.meriaudeau}@u-bourgogne.fr

Résumé – Les troubles d'humeur affectent de nombreuses personnes, la dépression étant la plus courante. Les méthodes avec la prospective d'aide aux cliniciens dans le diagnostic sont proposées ici, en fonction de la géométrie de l'expression du visage et de la parole. Les approches indépendantes du genre et dépendantes du genre ont été testées, pour différentes combinaisons de caractéristiques visuelles et audio. L'évaluation et la quantification des méthodes développées, pour plusieurs ensembles de paramètres, sont effectuées sur l'ensemble de données fournies par le challenge Emotion Audio / Visual. Un score de F1 de 71.3 % a été atteint pour détecter les individus signalant des scores élevés sur le BDI-II. La meilleure configuration du système comprenait un classificateur d'analyse discriminant pour les caractéristiques géométriques dans l'approche indépendante du genre.

Abstract – Mood disorders are burdening many individuals, with depression being the most common. Methods with the prospective of aiding clinicians in the diagnosis are being proposed hereby, based on facial expression geometry and speech. Gender independent and gender dependent approaches have been tested, for different combinations of visual and audio features. Evaluation of the algorithm, for several sets of parameters, was performed on the dataset provided by the Audio/Visual Emotion Challenge. F1-score of 71.3% was achieved for detecting individuals reporting high scores on BDI-II. The best system configuration included discriminant analysis classifier for geometrical features in the gender independent approach.

1 Introduction

De nos jours, la dépression est une maladie mentale qui touche beaucoup de monde. Un système destiné à l'estimation automatique de la dépression permet d'apporter des aides précieuses aux cliniciens et de mieux gérer les patients. D'après la littérature clinicienne, la géométrie faciale et la parole peuvent révéler des signes de dépression. Typiquement, les individus déprimés ont tendance à changer d'expression très lentement et prononcent des paroles plates avec des pauses étirées [1]. Dans cet article, nous présentons une méthode de la détection de dépression basée sur l'analyse de ces deux caractéristiques (géométrie faciale et parole) suivie d'une fusion des résultats. Dans la suite de l'article, nous traçons d'abord brièvement un état de l'art du domaine concerné, puis, la base de données utilisée sera décrite en Section 3. La méthodologie adoptée et les résultats expérimentaux seront respectivement présentés en Sections 4 et 5.

2 Etat de l'art

Il existe beaucoup de littérature dans le domaine de la détection automatique de la dépression et de la détermination du degré de cette dépression [2]. Cependant, les travaux qui concernent directement notre problématique de recherche viennent seulement d'apparaitre. Pour la détection de la dépression, deux types de caractéristiques sont souvent exploités : bas niveau et/ou haut niveau. Le haut niveau implique des informations pouvant être traduites en connaissances humaines et le bas niveau concerne principalement l'apprentissage automatique.

L'Outil OpenFace [3] développé par Baltrušaitis et al. permet d'extraire les point fiduciaires illustrés et numérotés dans les figures 1 et 2. Pampouchidou et al. [2] ont présenté une approche pour l'évaluation automatique de dépression basée sur les séries chronologiques de distances entre les repères et les caractéristiques issues de la parole. Une autre approche présentant des caractéristiques similaires est celle de Nasir et al. [4]. D'autres approches basées sur la géométrie dans le do-

maine de la reconnaissance de l'expression faciale (FER) comprennent celle de Joshi et al. [5], qui considérait la largeur et la hauteur des différentes caractéristiques faciales (yeux, bouche, etc.). Une autre approche de type FER proposé par Gacav et al. [6] a produit des points supplémentaires, en combinant X et Y de différents points. L'excentricité est une caractéristique qui a été proposée par Loconsole et al. [7], encore une fois pour FER.

3 Description de la base de données

La base de données utilisée pour tester notre méthode a été introduite à l'occasion des 3ème et 4ème AVEC [8] (Audio / Visual Emotion Challenge). Pour construire cette base de données, les participants volontaires ont été recrutés. Chaque individu possède une annotation correspondant au degré de dépression. Durant la phase d'acquisition, chaque participant doit réaliser plusieurs tâches, une webcam capture et enregistre les séquences vidéo. Pour notre travail, seulement une partie de la base AVEC, a été utilisée correspondant à deux tâches : a). FreeForm où les participants répondaient à des questions et b). NorthWind où les participants lisaient un passage à voix haute. Les données ont été réparties en 3 ensembles respectivement pour l'apprentissage, pour le développement et aussi pour le test. Sur un total de 300 participants, 200 possèdent une annotation de dépression.

Cette annotation correspond au score BDI-II (Beck Depression Inventory-II) de chaque participant. Ceux qui ont mis plus de temps pour accomplir une tâche, voient leur score augmenter. Habituellement, le score BDI-II est interprété de la manière suivante : a) [0-13] - dépression minimale, b) [14-18] - état dépressif, c) [19-28] - dépression légère, et d) [30-63] - dépression majeure.

L'objectif du sous-défi "AVEC Depression" était de prédire automatiquement le score BDI-II à partir d'une séquence vidéo. Ce score peut aussi être utilisé pour séparer deux groupes : dépression vs. non dépression.

4 Présentation de la méthodologie

Notre méthode consiste en une chaîne de traitement classique [2]. Le prétraitement constitue sa première étape, suivie par l'extraction des caractéristiques (voir les figures 1 et 2, et le tableau 1). Basé sur l'état de l'art, deux types de caractéristiques, respectivement provenant du vidéo et de l'audio, ont été exploités. L'ACP (Analyse en Composantes Principales) appliquée aux deux modalités permet de réduire la dimension des caractéristiques. Ensuite, les résultats de classification de deux modalités ont été fusionnés pour prendre une décision sur la présence des symptômes de dépression. 200 participants enregistrés sont divisés en deux classes : 96 avec symptômes de dépression significatif, et 104 sans.

4.1 Prétraitement et Extraction des caractéristiques

Le prétraitement de vidéo consiste à détecter et à repérer 68 points fiduciaires de deux dimensions en utilisant l'outil OpenFace [3]. Nous n'avons gardé que des séquences dont le logiciel localise bien les points significatifs pour la future analyse. Les erreurs de détection/localisation sont principalement dues à la mauvaise qualité de vidéo. Les enregistrements audio ont été sous échantillonnés à 16 kHz et les segments sans voix ont été enlevés des séquences temporelles. Ensuite, la fréquence fondamentale F0 a été estimée entre 70-450 Hz et normalisée entre 0 et 1 pour minimiser la dépendance au parleur.

Les caractéristiques de distance présentées dans [2] sont employées ici et améliorées avec des fonctionnalités supplémentaires issues des tavaux de [4, 5, 6, 7]. Toutes les fonctions de distance (antérieures et supplémentaires) sont présentées dans la figure 1, tandis que les nouvelles fonctionnalités sont présentées en figure 2.

L'ensemble de distances (voir la figure 1) est calculé image par image pour chaque séquence vidéo. Toutes ces distances ont été normalisées par la largeur du visage (points 1 et 17 sur la figure 1) pour prendre en compte le changement de distances dûaux mouvements de la tête.

Ensuite plusieurs paramètres statistiques ont été calculés à partir de la mesure d'une distance extraite d'une séquence vidéo complète. Ces paramètres sélectionnés sont au nombre de douze : moyenne, médiane, mode, étendue, écart type, variance, asymétrie, kurtosis, énergie, entropie, corrélation, étendue d'interquartile, max, min, mad, fréquence moyenne et puissance de bande. La largeur de l'œil droit et gauche, ainsi que la largeur et la hauteur de la bouche illustrées dans la figure 2, ont également été considérées de manière chronologique pour une extraction supplémentaire des caractéristiques.

Nous avons aussi additionné un autre ensemble des caractéristiques basé sur les activités des points fiduciaires en termes de déplacement, de vitesse et d'accélération. Ces paramètres sont mesurés à l'aide d'une fenêtre temporelle glissante. Tous les points sont regroupés en six zones illustrées sur la figure 1 : sourcils gauche/droite, yeux gauche/droit, bouche et visage entier. Pour chaque zone, on moyenne les valeurs de déplacement, de vitesse et d'accélération sur ces points encadrés avant de calculer les paramètres statistiques correspondants. Les caractéristiques de mouvement de référence ont été calculées en plus pour les points projetés indiqués dans la figure 2.

Les caractéristiques d'excentricité présentées dans [7] ont été calculées, selon la formule suivante :

$$e = \frac{\sqrt{a^2 - b^2}}{a} \tag{1}$$

où $a = \frac{B_{Mx} - A_{Mx}}{2}$ et $b = A_{My} - U_{m1y}$, pour l'exemple de l'excentricité de la bouche supérieure, comme indiqué dans la figure 2. Pour la bouche inférieure U_{m1y} est remplacé par D_{m2}, et l'excentricité pour le reste des traits du visage est calculée de la même manière. Pour le cas des yeux, les points centraux

ne sont pas présents, ils ont été estimés avec une interpolation cubique basée sur les repères voisins. Au total nous avons 1751 caractéristiques de la géométrie faciale.

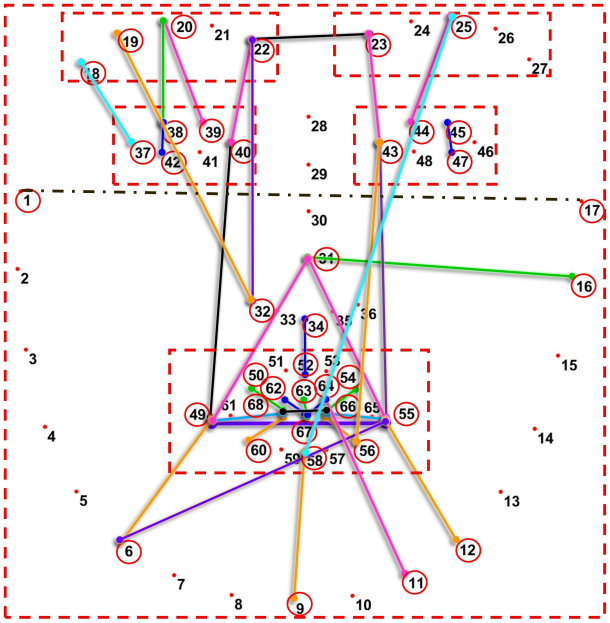

FIGURE 1 – Mesure des distances entre les points géométriques ; les rectangles rouges correspondent aux zones considérées pour la mesure des activités de points

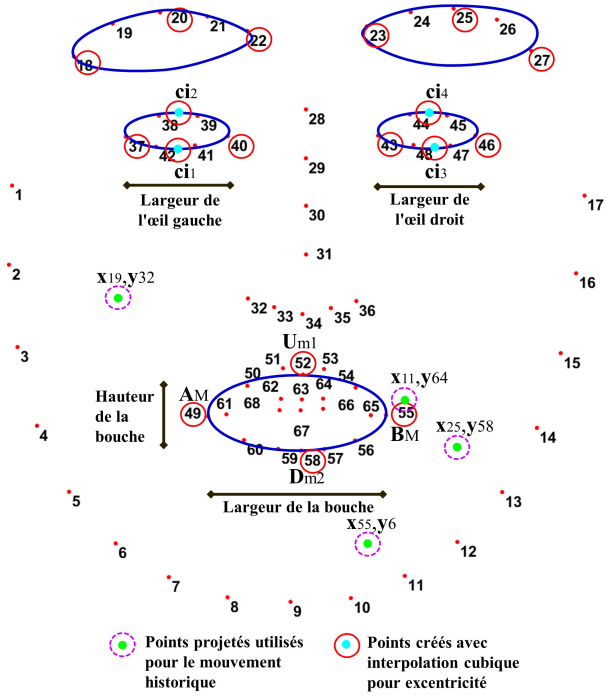

FIGURE 2 – Caractéristiques géométriques supplémentaires. Les lignes noires correspondent aux caractéristiques largeur / hauteur. Les ellipsoïdes bleus correspondent aux caractéristiques d'excentricité.

Pour extraire les caractéristiques audio, nous avons utilisé le logiciel en libre service COVAREP (V.1.4.2) [9]. Les descripteurs de bas niveau, extraits de la même manière que [2], sur l'ensemble d'enregistrement audio avec un intervalle de 10 ms, sont résumés dans le tableau 1.

TABLE 1 – Descripteurs statistiques calculés avec des caractéristiques audio [2])

Descripteurs de bas niveau
normalized F_0
delta F_0, delta-delta F_0
NAQ, QOQ, H1H2, PSP, MDQ, peakSlope, Rd, Rd conf
MCEP 0-24
delta MCEP 0-24, delta-delta MCEP 0-24
HMPDM 1-24, HMPDD 1-12
Formants 1-3

4.2 Fusion et Classification

Nous avons testé deux schémas de fusion : la fusion au niveau de caractéristiques et la fusion au niveau de décision. Le premier concatène les vecteurs de caractéristiques de deux modalités vidéo/audio en un seul avant de passer à l'étape de la classification. Pour le deuxième, on a deux classifieurs (un pour chaque modalité). La décision finale est faite par une simple opération logique entre les deux résultats préliminaires. Les classifications ont été effectuées avec deux modes : Groupe Mixte (GM) et Groupe Non-Mixte (GNM). C'est-à-dire que l'on mélange ou non les deux groupes d'individus de genre différent : homme et femme.

5 Résultats expérimentaux

Nous avons expérimenté plusieurs configurations de la méthode proposée. Par exemple, la distance entre les points fiduciaires a été calculée avec la distance Euclidienne et la distance Minkowski et la fenêtre temporelle utilisée pour l'extraction des activités de mouvement varie de 15 à 60 images, ceci correspond à un enregistrement de 0.5 s à 2 s (avec une vitesse de 30 images par seconde).

Pour les caractéristiques audio, deux ensembles de descripteurs ont été testés : a) descripteurs statistiques de bas niveau 1 et b) les dix premiers coefficients de la DCT (Discrete Cosine Transform) sur les descripteurs de bas niveau. L'ACP a aussi été appliquée avec plusieurs variantes en gardant 50, 60, 70, 80, 90, 100, 150, 160, 170, 180, et 199 composantes. Nous avons testé deux type de classifieurs : le plus proche voisin et analyse discriminante. La meilleure configuration est la suivante : la distance Euclidienne, 170 composantes gardées pour l'ACP pour les deux modalités vidéo et audio, avec le classifieur de type Analyse discriminante.

Les résultats sont résumés dans les tableaux 2, 3,4, et 5. Les

tableaux 2 et 3 affichent le score F1. Le tableau 3 décrit les tests réalisés avec la modalité vidéo en fonction de la taille de la fenêtre temporelle. Le résultat optimal correspond à une fenêtre de 25 images (830ms) avec APC 170. Les différents schémas de fusion ont été analysés à l'aide du tableau 4.

TABLE 2 – Résultats de la modalité Audio (Score F1)

	DCT	**Functionals**
Groupe Mixte (GM)	0.588	0.626
Groupe Non-Mixte(GNM)	**0.641**	0.545

TABLE 3 – Résultats de la modalité Vidéo (Score F1)

Fenêtre / APC	**15**	**20**	**25**	**30**	**60**
160	66.4	68.2	68.7	66.4	65.2
170	65.7	68.1	**71.3**	67.6	65.7

TABLE 4 – Résultats avec différents schémas de fusion

	Précision	**Rappel**	**F1**
Vidéo GM	0.564	0.969	**0.713**
Vidéo GNM	**0.625**	0.052	0.96
Audio GM	0.556	0.625	0.588
Audio GNM	0.6	0.688	0.641
Concaténation Fusion GM	0.475	0.604	0.532
Vidéo GM ET Audio GM	0.606	0.594	0.6
Vidéo GM OU Audio GNM	0.542	**1**	0.703

TABLE 5 – Matrice de confusion de classification (prédiction) obtenue avec 200 individus (104 non dépressifs et 96 dépressifs) avec Vidéo GM

Prédiction / Self	**Non dépressifs**	**Dépressif**
Non dépressifs	**32 (30.8%)**	72 (69.2%)
Dépressif	3 (3.1%)	**93 (96.9%)**

6 Conclusions

Cet article présente nos travaux de détection de la dépression en analysant deux modalités : la géométrie faciale (vidéo) et la parole (audio). L'objectif est de développer un système d'aide au diagnostic dans le contexte de télé-psychologiques applications et de programmes d'enquête (prévention) sur des larges populations. La performance optimale a été obtenue sur la modalité vidéo avec le mode "Groupe Mixte". Concernant la modalité audio, le mode "Groupe Non-Mixte" fonctionne le mieux. La fusion au niveau de décision a donné de meilleures performances en termes de rappel, car il atteint 100 % d'identification correcte sur des personnes déprimées.

Malgré une précision de détection qui n'est pas assez bonne pour les individus sans symptômes significatifs, les résultats obtenus par ce travail sont encourageants. Nous envisageons d'améliorer la méthode en termes de spécificité pour mieux viser la fausse détection en vue d'un système autonome de détection automatique de la dépression. Plusieurs pistes seront exploitées : a) améliorer la chaîne de traitement, b) optimiser la détection des caractéristiques négativement corrélées avec la dépression et c) fusionner d'autres informations complémentaires (i.e. des tests psychométriques et des signaux physiologiques). Le but final est d'établir un système multi-facteurs fiable et robuste.

7 Acknowledgments

Anastasia PAMPOUCHIDOU bénéficie d'une bourse de la fondation "Greek State Scholarships Foundation" - I.K.Y).

Références

[1] A. Pampouchidou et al., "Automatic assessment of depression based on visual cues : A systematic review," *IEEE Transactions on Affective Computing*, 2017.

[2] A. Pampouchidou et al., "Facial geometry and speech analysis for depression detection," in *39th Annual Intern. Conf. of the IEEE EMBS*, 2017.

[3] T. Baltrušaitis et al., "OpenFace : An Open Source Facial Behavior Analysis Toolkit," in *IEEE Winter Conf. on Applications of Computer Vision*, 2016, pp. 1–10.

[4] Md Nasir et al., "Multimodal and Multiresolution Depression Detection from Speech and Facial Landmark Features," in *6th Intern. Workshop on Audio/Visual Emotion Challenge*. 2016, pp. 43–50, ACM.

[5] A. Joshi et al., "Predicting active facial expressivity in people with parkinson's disease," in *9th ACM Intern. Conf. on PErvasive Technologies Related to Assistive Environments*. ACM, 2016, p. 13.

[6] C. Gacav et al., "Greedy search for descriptive spatial face features," *arXiv preprint :1701.01879*, 2017.

[7] C. Loconsole et al., "Real-time emotion recognition novel method for geometrical facial features extraction," in *Intern. Conf. on Computer Vision Theory and Applications*. IEEE, 2014, vol. 1, pp. 378–385.

[8] M. Valstar et al., "AVEC 2014 : 3D Dimensional Affect and Depression Recognition Challenge," in *4th Intern. Workshop on Audio/Visual Emotion Challenge*. ACM, 2014, pp. 3–10.

[9] G. Degottex et al., "COVAREP - A Collaborative Voice Analysis Repository for Speech Technologies," in *IEEE Intern. Conf. on Acoustics, Speech and Signal Processing*, 2014, pp. 960–964.