Analyse multimodale d'émotions utilisant l'audio, la vidéo, la transcription et des signaux physiologiques

Soëlie Lerch, Patrice Bellot, Emmanuel Bruno, Elisabeth Murisasco

Analyse multimodale d'émotions utilisant l'audio, la vidéo, la transcription et des signaux physiologiques

Soëlie Lerch*
soelie.lerch@lis-lab.fr
Aix Marseille Univ, Université de Toulon, CNRS, LIS
Toulon, France

Patrice Bellot
patrice.bellot@univ-amu.fr
Aix Marseille Univ, Université de Toulon, CNRS, LIS
Marseille, France

Emmanuel Bruno
Elisabeth Murisasco
prenom.nom@lis-lab.fr
Aix Marseille Univ, Université de Toulon, CNRS, LIS
Toulon, France

ABSTRACT

Nous étudions l'impact de l'ajout de plusieurs modalités à la transcription de la parole pour la prédiction d'émotions. Ce travail a pour objectif de contribuer aux thérapies pour l'éducation émotionnelle des adolescents TSA (Trouble du Spectre Autistique) à partir d'une recommandation automatique de passages de séries TV.

KEYWORDS

Analyse multimodale d'émotions, Signaux physiologiques, Audio, video, Transcription, Trouble du spectre autistique

1 CONTEXTE ET MOTIVATION

La détection d'émotions quand plusieurs individus communiquent dans un langage verbal ou non verbal permet de recevoir des messages implicites sur la compréhension de la conversation. Dans différents travaux, l'émotion est principalement représentée sous deux formes : des noms de catégories, comme celles d'Ekman [9] ou selon plusieurs dimensions [18] : la valence (degré de plaisir ou de déplaisir), l'activation (intensité de l'état d'esprit), la dominance (niveau de contrôle sur l'émotion). Par exemple, la joie a une valence très positive et une activation forte avec une dominance moyenne à forte. Des travaux se sont focalisés sur l'analyse d'émotions exprimées dans des dialogues humains en ne considérant que la modalité textuelle [7] sur des dialogues courts dont certains sont ambigus. Des indices tels que les expressions faciales, la prosodie, des postures ou des signaux physiologiques peuvent permettre de mieux comprendre l'état émotionnel de la personne.

Nous voulons mesurer le degré d'impact de l'ajout de signaux physiologiques dans la prédiction d'émotions multimodales. Ces signaux sont enregistrés sur une personne qui regarde un film afin de détecter les émotions véhiculées par ce stimulis. Nous nous intéressons plus spécifiquement aux troubles de la compréhension et de l'expression des émotions des personnes TSA (Trouble du Spectre Autistique). Notre but est d'étudier leurs émotions pour contribuer aux thérapies cognitives qui y sont dédiées. Actuellement, ces thérapies sont basées sur des travaux pratiques à partir de scènes visuelles ou de dessins animées [21] pour l'apprentissage de reconnaissance d'expressions faciales. Nous proposons d'aider le thérapeute à identifier des scènes de séries TV correspondantes aux émotions pour améliorer l'engagement et la participation du jeune TSA. Pour cela, il faut créer une modélisation des émotions applicables aux séquences vidéo. Cette modélisation est le lien entre un modèle humain (représenté par des descripteurs physiologiques et un modèle linguistique) et les types d'émotions. Elle permettra d'aider à configurer des agents virtuels pour que leur expression soit perçue correctement par les personnes TSA. Pour mieux cerner la difficulté de compréhension des émotions des TSA, nous comparerons via des approches informatiques du traitement automatique des langues et de fouille de données, les émotions ressenties de deux groupes adolescents neurotypiques (non autistes) et TSA. Nous étudierons l'émotion ressentie par le sujet et l'émotion induite (celle présente dans la vidéo) dans une démarche similaire à [17].

L'état de l'art montre que les signaux physiologiques permettent la prédiction de l'état émotionnel d'une personne [8]. Différents travaux utilisent l'électrocardiogramme (ECG), certains classifient l'activation et la valence [19] grâce à l'usage de réseaux de neurones convolutifs CNN [15]; d'autres [6] étudient la prédiction d'émotions lors du visionnement de films de deux personnes selon leur lien (inconnus, amis, couple) avec une représentation de l'émotion par catégorie; en plus de l'ECG, la respiration [2] est utilisée pour prédire l'émotion représentée par l'activation, la valence et la dominance parmi deux émotions dans le cas d'une écoute musicale. D'autres enfin [14][13] prédisent des catégories d'émotion uniquement selon la respiration (rythme, intensité) durant le visionnage de courtes vidéos.

Pour la détection des mouvements oculaires, des travaux [22] prédisent la valence et l'activation en associant les descripteurs liés à la focalisation du regard à ceux de l'image selon plusieurs classifieurs. D'autres [16] utilisent des réseaux adverses génératifs Wasserstein (WGAN) [3] pour augmenter le jeu de données d'entraînement des mouvements oculaires et de l'électroencéphalogramme (EEG) pour la prédiction de catégories d'émotion. Certains [5] considèrent, en plus des mouvements oculaires, des signaux physiologiques, étudiés séparément, pour une comparaison entre des personnes TSA et neurotypiques.

Nous présentons dans la section 2, le protocole de notre expérimentation en cours de réalisation, de la collecte des descripteurs à la prédiction d'émotions à partir de la vidéo (image, audio, transcription) et des signaux physiologiques (ECG, respiration, mouvements oculaires) sur les personnes TSA et non TSA.

WACAI 2020 '20, June 03–05, 2020, Île d'Oléron, France

ACM ISBN 978-1-4503-XXXX-X/18/06…$15.00
https://doi.org/10.1145/1122445.1122456

2 PROTOCOLE D'EXPÉRIMENTATION

Lors de la mesure de la réaction émotionnelle des personnes TSA, hypersensibles, l'usage de capteurs ne doit pas engendrer un niveau de stress trop important pour que les mesures des signaux physiologiques puissent être exploitables. Nous utilisons des capteurs non intrusifs tels que la montre Apple Watch pour le relevé de l'ECG, l'Eye tracker Tobii pro nano pour la détection de la focalisation du regard et de la dimension de la pupille ainsi qu'un capteur à ultrasons [4] pour une détection non intrusive de la respiration grâce à la perturbation de l'air.

Le sujet est installé à environ cinquante centimètres de l'écran d'un ordinateur portable. Il porte à son poignet la montre. Le détecteur de mouvements oculaires est placé au bas de l'écran, le capteur à ultra-sons au-dessus de l'écran. Le sujet visualise des vidéos provenant de la base de donnée LIRIS-accede (https://liris-accede.ec-lyon.fr/database.php) pour l'entraînement, c'est une collection de films et de signaux physiologiques avec des annotations émotionnelles basées sur la valence et l'activation, déjà utilisée pour la prédiction d'émotions [17] ; une deuxième jeu de données sera élaboré à partir d'un corpus de séries télévisées annoté en catégories d'émotions ainsi que la valence et l'éveil.

Les données physiologiques sont filtrées et associées à un horodatage pour les synchroniser avec la vidéo. Nous utilisons les données d'entrées suivantes : signaux physiologiques, expressions faciales, audio, image, transcription, questionnaire sur l'état émotionnel du sujet. Les données récoltées échantillonnées correspondent à une seconde d'enregistrement. Cependant, le signal généré par le capteur de souffle est perturbé par les mouvements du sujet. Nous envisageons deux façons d'exploiter cette particularité : filtrer le mouvement du corps à l'aide d'un apprentissage dédié, considérer le mouvement comme un indice pour la détection d'émotions. Si besoin, nous augmenterons les données avec des réseaux WGAN.

Pour les expressions faciales, nous utilisons la librairie OpenFace de Python. [1] [1]. Les points FACS (facial action coding system)[10] permettent de détecter les émotions grâce aux distances entre les points du muscle en repos et ceux du muscle en mouvement. Pour l'audio, nous utilisons OpenSmile [2] [11] pour extraire les descripteurs d'une musique ou de la parole. Pour les images, nous utilisons des CNN pour sélectionner les descripteurs.

Pour l'apprentissage, nous pouvons utiliser un classifieur binaire par émotion mais aussi deux classifieurs qui déterminent la valence et l'activation et en déduire l'émotion. Plusieurs apprentissages supervisés sont expérimentés [12]. Enfin, nous étudions le lien entre l'émotion induite et l'émotion de la personne analysée (neurotypique ou TSA). Nous déterminons le lien entre ces émotions et les descripteurs qui y sont associés pour les deux types de sujet.

3 CONCLUSION

L'objectif de notre étude en cours est de déterminer à quel point des indices physiologiques améliorent les analyses dans le cadre de l'analyse de sentiment[20] ou d'émotion[7]. Nous souhaitons utiliser l'analyse multimodale des émotions pour créer un moteur de recherche de scènes inductives d'émotions pour la recommandation de contenus en fonction de ce que le téléspectateur souhaite ressentir plutôt que par genre. Cette étude contribue également aux thérapies cognitives avec la configuration d'agents virtuels pour l'éducation émotionnelle de jeunes TSA.

[1] http://cmusatyalab.github.io/ openface/
[2] https://www.audeering.com/ opensmile/

REFERENCES

[1] Brandon Amos, Bartosz Ludwiczuk, Mahadev Satyanarayanan, et al. 2016. Openface: A general-purpose face recognition library with mobile applications. *CMU School of Computer Science* 6 (2016), 2.

[2] K Anuharshini, M Sivaranjani, M Sowmiya, V Mahesh, and B Geethanjali. 2019. Analyzing the Music Perception Based on Physiological Signals. In *2019 5th International Conference on Advanced Computing & Communication Systems (ICACCS)*. IEEE, IEEE, Coimbatore, India, 411–416.

[3] Martin Arjovsky, Soumith Chintala, and Léon Bottou. 2017. Wasserstein generative adversarial networks. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*. Sydney, Australia, 214–223.

[4] Philippe Arlotto, Michel Grimaldi, Roomila Naeck, and Jean-Marc Ginoux. 2014. An ultrasonic contactless sensor for breathing monitoring. *Sensors* 14, 8 (2014), 15371–15386.

[5] Pradeep Raj Krishnappa Babu and Uttama Lahiri. 2019. Classification approach for understanding implications of emotions using eye-gaze. *Journal of Ambient Intelligence and Humanized Computing* (2019), 1–13.

[6] Andrea Bizzego, Atiqah Azhari, Nicola Campostrini, Anna Truzzi, Li Ying Ng, Giulio Gabrieli, Marc H Bornstein, Peipei Setoh, and Gianluca Esposito. 2020. Strangers, Friends, and Lovers Show Different Physiological Synchrony in Different Emotional States. *Behavioral Sciences* 10, 1 (2020), 11.

[7] Ankush Chatterjee, Kedhar Nath Narahari, Meghana Joshi, and Puneet Agrawal. 2019. SemEval-2019 task 3: EmoContext contextual emotion detection in text. In *Proceedings of the 13th International Workshop on Semantic Evaluation*. 39–48.

[8] JA Domínguez-Jiménez, KC Campo-Landines, JC Martínez-Santos, EJ Delahoz, and SH Contreras-Ortiz. 2020. A machine learning model for emotion recognition from physiological signals. *Biomedical Signal Processing and Control* 55 (2020), 101646.

[9] Paul Ekman. 1992. An argument for basic emotions. *Cognition & emotion* 6, 3-4 (1992), 169–200.

[10] Rosenberg Ekman. 1997. *What the face reveals: Basic and applied studies of spontaneous expression using the Facial Action Coding System (FACS)*. Oxford University Press, USA.

[11] Florian Eyben, Felix Weninger, Florian Gross, and Björn Schuller. 2013. Recent developments in opensmile, the munich open-source multimedia feature extractor. In *Proceedings of the 21st ACM international conference on Multimedia*. Association for Computing Machinery, Barcelona, Spain, 835–838.

[12] Ian Goodfellow, Yoshua Bengio, and Courville Aaron. 2016. *Deep learning*. MIT Press.

[13] Carolina Gouveia, Ana Tomé, Filipa Barros, Sandra C Soares, José Vieira, and Pedro Pinho. 2020. Study on the usage feasibility of continuous-wave radar for emotion recognition. *Biomedical Signal Processing and Control* 58 (2020), 101835.

[14] Rabab A Hameed, Mohannad K Sabir, Mohammed A Fadhel, Omran Al-Shamma, and Laith Alzubaidi. 2019. Human emotion classification based on respiration signal. In *Proceedings of the International Conference on Information and Communication Technology*. Baghdad, Iraq, 239–245.

[15] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. 1998. Gradient-based learning applied to document recognition. *Proc. IEEE* 86, 11 (1998), 2278–2324.

[16] Yun Luo, Li-Zhen Zhu, and Bao-Liang Lu. 2019. A GAN-Based Data Augmentation Method for Multimodal Emotion Recognition. In *International Symposium on Neural Networks*. Springer, Springer, Cham, 141–150.

[17] Michal Muszynski, Leimin Tian, Catherine Lai, Johanna Moore, Theodoros Kostoulas, Patrizia Lombardo, Thierry Pun, and Guillaume Chanel. 2019. Recognizing induced emotions of movie audiences from multimodal information. *IEEE Transactions on Affective Computing* (2019).

[18] J.A. Russel. 1980. A circumplex model of affect. *Journal of Personality and Social Psychology* 39, 6 (1980), 1161–1178.

[19] Pritam Sarkar and Ali Etemad. 2019. Self-supervised Learning for ECG-based Emotion Recognition. *arXiv preprint arXiv:1910.07497* (2019).

[20] Lerch Soëlie, Bellot Patrice, Bruno Emmanuel, and Murisasco Elisabeth. 2019. Influence des lexiques d'émotions et de sentiments sur l'analyse de sentiments-Application à des critiques de livres. In *COnférence en Recherche d'Informations et Applications-CORIA 2019, 16th French Information Retrieval Conference*. Villeurbanne, France.

[21] A. T. Wieckowski and S. W. White. 2020. Attention Modification to Attenuate Facial Emotion Recognition Deficits in Children with Autism: A Pilot Study. *Journal of Autism and Developmental Disorders* 50, 1 (2020), 30–41.

[22] Atiyeh Yaghoubiy, Seyed Kamaledin Setarehdan, and Keivan Maghooli. 2019. Emotion Extraction from Video Fragments using Gaze Tracking and AdaBoost Classifier. *Majlesi Journal of Electrical Engineering* 13, 2 (2019), 67–81.