EMOLIS App et Dataset pour suggérer des dessins animés proches émotionnellement

Soëlie Lerch, Patrice Bellot, Emmanuel Bruno, Elisabeth Murisasco

EMOLIS App et Dataset pour suggérer des dessins animés proches émotionnellement

Soëlie Lerch[1], Patrice Bellot[2,‡], Emmanuel Bruno[1] and Elisabeth Murisasco[1,†]

[2] *LIS UMR 7020 CNRS / AMU / UTLN, Aix Marseille Université Campus de Saint Jérôme, 13 3997 Marseille Cedex 20, France*

[1] *LIS UMR 7020 CNRS / AMU / UTLN, Université de Toulon Campus de La Garde, 83041 Toulon Cedex 9, France*

Abstract

Nous proposons EMOLIS Dataset, un jeu de données d'annotations d'émotions et de signaux physiologiques enregistrés sur des spectateurs (respiration, électrocardiogramme, mouvements oculaires) regardant des dessins animés de Walt Disney. EMOLIS Dataset ne contient ni les vidéos, ni les transcritions mais des pointeurs temporels vers les sources. Il a été créé pour la conception d'EMOLIS App, une application qui affiche les émotions en temps réel d'une vidéo et suggère des vidéos similaires à un spectateur à partir du ressenti émotionnel. Pour cela, nous estimons une distance émotionnelle entre des vidéos en utilisant des représentations multimodales des vidéos et des descripteurs de signaux physiologiques associés. Ces résultats personnalisés peuvent être utilisés dans le cadre de thérapies cognitives sur la conscience des émotions ressenties. Le jeu de données est conçu pour être accessible à toutes les audiences et aux personnes atteintes de Troubles du Spectre Autistique (TSA) qui peuvent avoir des difficultés à reconnaître les émotions ressenties.

Keywords

Analyse des émotions, Suggestion de vidéos, Suggestion interactive, Trouble du spectre autistique

1. Introduction

Cet article s'inscrit dans le cadre de l'analyse automatisée des émotions et cible à la fois les émotions émises et le ressenti des spectateurs. Le distinguo entre ces deux types d'émotions, émotions perçues et émotions ressenties, est important comme dans tout processus de communication où l'intention de l'émetteur est réceptionnée par un destinataire plus ou moins fidèlement. Les scènes d'un dessin animé comportent des signaux visuels, sonores et sémantiques qui communiquent des émotions qui sont réceptionnées par des spectateurs et qui engendrent, à différents degrés, des réactions physiologiques et une prise de conscience du ressenti. Lorsque l'on s'intéresse à la qualification automatisée des émotions véhiculées par un contenu, par exemple dans un objectif de classification des contenus ou pour rechercher des contenus émotionnellement similaires, se pose alors la question de savoir jusqu'à quel point l'on peut se contenter de considérer seulement les contenus, indépendamment des ressentis. Pour répondre à cette question, nous pouvons commencer par intégrer dans un système de classification différents modules d'analyse d'émotions de contenus vidéo ciblant les expressions

COnférence en Recherche d'Informations et Applications-CORIA 2024, 19th French Information Retrieval Conference, La Rochelle, France

soelie.lerch@lis-lab.fr (S. Lerch); patrice.bellot@univ-amu.fr (P. Bellot); emmanuel.bruno@lis-lab.fr (E. Bruno); elisabeth.murisasco@lis-lab.fr (E. Murisasco)

faciales des personnages, la voix mais aussi le contenu sémantique des dialogues. Cette intégration de différents prédicteurs d'émotions permet d'identifier des émotions dominantes pour les comparer à une vérité terrain déterminée selon une annotation humaine. S'il est d'usage d'étudier l'accord inter-annotateurs dans toute procédure de modélisation par apprentissage supervisé et d'évaluation selon une vérité terrain, ce point est ici encore plus crucial puisque les émotions sont avant tout subjectives. L'existence même d'une telle vérité terrain pose en effet question tant le ressenti des spectateurs est variable selon leur sensibilité et leur vécu et tant la qualification des émotions dépend de la capacité à verbaliser les ressentis selon des normes socioculturelles. Pour apprendre un modèle de classification, une alternative à une annotation manuelle peut être alors de considérer non pas les émotions nommées par les spectateurs mais les signaux physiologiques eux-mêmes: rythme cardiaque, respiration, suivi visuel…

Dans cet article, nous présentons un état de l'art et une discussion des bases de données existantes (section 2). Nous décrivons EMOLIS Dataset un jeu de données d'annotations d'émotions et de signaux physiologiques enregistrés sur des spectateurs (respiration, électrocardiogramme, mouvements oculaires) regardant des dessins animés de Walt Disney (section 3). EMOLIS Dataset ne contient ni les vidéos, ni les transcriptions mais des pointeurs temporels vers les sources. Nous introduisons une application EMOLIS App qui permet d'afficher les émotions en temps réel grâce à un modèle multimodal appris et aussi la suggestion de vidéos estimées similaires d'un point de vue émotionnel à partir des représentations d'un réseau de neurones multimodal (texte, audio, vidéo) et les descripteurs de signaux physiologiques (section 4).

2. Etat de l'art et jeux de données existants

Nous recherchons des vidéos courtes accessibles à tout public qui soient porteuses des émotions d'Ekman [1] : joie, colère, tristesse, peur, dégoût, surprise et qui contiennent des dialogues. Nous privilégions les dessins animés pour cibler la reconnaissance d'émotions sur un public jeune et sur des personnes TSA (Trouble du Spectre Autistique) qui peuvent avoir des difficultés à reconnaître les émotions qu'ils ressentent. Nous choisissons des vidéos de type dessins animés car les scènes sont généralement moins riches en objets perturbateurs, avec des visages aux traits simplifiés [2] ainsi qu'une vitesse de variation graphique moindre : des variations rapides peuvent interférer avec la compréhension des émotions perçues [3]. Les thérapies cognitives se concentrent sur la reconnaissance des émotions en apprenant à partir d'images, de vidéos et de chansons [1]. Plusieurs programmes d'entraînement sur ordinateur ont été proposés [4] mais les corpus contiennent peu de vidéos et sont coûteux à construire. Notre idée est de proposer une méthode qui facilite la création de corpus annotés par un modèle de prédiction d'émotions. Pour les dessins animés, nous n'avons trouvé qu'un seul corpus déjà constitué, mais contenant uniquement des expressions faciales des personnages de Tom et Jerry [5].

Iemocap [6] contient douze heures de conversations en anglais filmées entre 10 acteurs ainsi que les sous-titres par tours de parole. Les émotions présentes sont celles d'Ekman, la frustration, l'excitation, l'état neutre et une classe qui englobe d'autres émotions que celles proposées. La représentation en trois dimensions (valence, activation, dominance) [7] est aussi annotée. La valence correspond à la polarité de l'émotion (positive, négative), l'activation à l'intensité de

[1] voir par exemple https://modelmekids.com/autism-emotions/.

l'émotion et la dominance au degré de contrôle de l'émotion. Elle a été construite pour détecter les émotions perçues par le spectateur non les émotions ressenties.

Liris-Accede [8] contient 9800 extraits de films pour un total de 26 heures 57 minutes. Le contenu est varié afin d'exprimer différentes émotions chez le spectateur. La majorité de ces extraits est en anglais. Quelques scènes de dessins animés sont présentes avec peu de dialogues. L'annotation concerne la valence et l'activation. Dans le cadre de notre étude, nous souhaitons que les émotions soient liées à un contexte court afin que nous sachions par quel évènement précis l'émotion a été déclenchée. Les vidéos de Liris-Accede sont trop longues et le contenu de certaines d'entre elles ne sont pas adaptés pour les enfants pour les émotions peur et dégoût.

Mahnob-hci [9] contient 155 scènes de films associées à des signaux physiologiques tels que l'électrocardiogramme, la température et l'activité électrique de la peau, les mouvements oculaires et les expressions faciales des spectateurs, annotées selon les émotions d'Ekman, l'état neutre, la valence et l'activation et la dominance. Amigos [10] contient des vidéos courtes stimuli et des signaux physiologiques tels que l'électrocardiogramme, l'électroencéphalogramme, la conductance électrique de la peau, annotées avec la valence, l'activation, la dominance, les émotions d'Ekman et l'état neutre. Bien que les annotations de ces deux derniers corpus correspondent à nos attentes, ils ne sont pas adaptés à un jeune public ou à un public hyper sensible, par exemple en situation de handicap (le film d'horreur Shining).

MELD est issu de la série comique Friends [11] en anglais, accesible à des adolescents,. Les acteurs expriment les six émotions d'Ekman et l'état neutre et la polarité de l'émotion (positive, neutre, négative). La joie est l'émotion principalement ressentie, les autres sont peu présentes

3. EMOLIS Dataset

3.1. Modalités extraites de la vidéo

Faute de jeu de données disponible, nous en avons créé un nouveau. EMOLIS Dataset est composé de 62 scènes de une à quatre minutes extraites de huit longs métrages de W. Disney (Table 1), disponibles dans une cinquantaine de langues, incluant les modalités texte, image et audio. Pour chaque film, les scènes sont délimitées par un changement de décor ou de sujet de conversation entre les personnages. C'est pourquoi elles sont courtes. Ces scènes présentent un panel varié des émotions d'Ekman. Nous avons manuellement segmenté les scènes en séquences : une séquence correspond à un tour de parole, ou à un silence d'au moins une seconde. Cela permet de prédire les émotions en temps réel, par personnage et par moment de silence (expressions faciales). La durée totale du corpus EMOLIS Dataset est de 2h 8mn 38s.

Les sous-titres des scènes ont été obtenues à partir de sites en ligne, [2]. Les scripts que l'on va partager vont permettre d'aligner le texte avec la vidéo. Ces sous-titres sont une succession de dialogues que nous avons divisés en tours de parole. Pour chaque film, nous avons construit un fichier CSV contenant une ligne par séquence (tour de parole ou de silence), avec le numéro de la scène, le numéro du tour de parole, le nom du personnage qui parle et le début et la fin des marqueurs de la vidéo. La table 4 montre un exemple d'un fichier de données qui permet d'aligner les vidéos (table 2) et les sous-titres (table 3). A chaque instant t (Table 2), une séquence

[2] voir par exemple https://fr.my-subs.co/film-versions-4623-frozen-subtitles

Table 1
Les 62 scènes issues des dessins-animés Disney

Titre du dessin animé	nombre de scènes	nombre de séquences	durée
Planète au trésor	7	208	17:20
La reine des neiges	7	198	16:18
Mulan	9	207	17:20
Pocahontas	7	167	14:27
Rebelle	7	230	19:27
Vaiana	8	157	12:57
Vice-Versa	8	167	13:51
Zootopie	9	203	16:58
Total	62	1537	2:08:38

Table 2
Représentation temporelle des tours de paroles _ représente des silences et les lettres des phrases.

Personnage	t_1	t_2	t_3	t_4	t_5	t_6	t_7	t_8	t_9	t_{10}				
Elsa	A	B	C	D	_	_	_	_	E	F	_	G	_	H
Anna	_	_	V	W	X	Y	_	_	_	Z	_	_	_	_

Table 3
Fichier sous-titre à aligner avec la vidéo

ligne #	Personnage	p_1	p_2	p_3	p_4	t_{debut}	t_{fin}
1	Elsa	A	B	C	D	t_1	t_4
2	Anna	V	W	X	Y	t_3	t_6
3	Elsa	E	F			t_9	t_{10}
4	Anna	Z				t_{10}	t_{10}

est prononcée ou silencieuse et des personnages peuvent parler en même temps (instant t_3).

EMOLIS Dataset est disponible sur Zenodo.org avec les scripts pour extraire et formater les données ainsi que pour aligner les vidéos et les sous-titres.

3.2. Récolte des signaux physiologiques

EMOLIS Dataset contient aussi des signaux physiologiques qui peuvent être utilisés pour personnaliser la suggestion de vidéos. Nous avons récolté la respiration, l'électrocardiogramme et les mouvements oculaires de personnes qui regardent les scènes. (voir Figure 1).

La récolte des signaux se déroule de la manière suivante. Le spectateur est assis devant un ordinateur qui diffuse le film et enregistre les mouvements oculaires via le détecteur Tobii Pro Nano placé en bas de l'écran, et les expressions faciales via la caméra intégrée. Une Apple Watch enregistre l'électrocardiogramme. Enfin, un analyseur de souffle [12] positionné autour de l'écran, détecte la respiration via des ultra-sons par quatre micros et les enregistre sur un ordinateur. Nous calculons la représentation domaine de fréquences à partir des ultra-sons.

Table 4
Données pour aligner les sous-titres et la vidéo. La ligne 1 indique qu'Elsa parle de l'instant t_1 à t_4. Le sous-titre est à la ligne 1 de la table 3. La Ligne 3 représente un silence (personne ne parle de l'instant t_7 à t_8). Les lignes 1-2 et 4-5 montrent qu'Elsa et Anna parlent en même temps aux instants t_3, t_4 et t_{10}.

	Scène	Séquence	Personnage	Début	Fin
1	6	1	Elsa	$t1$	$t4$
2	6	2	Anna	$t3$	$t6$
3	6	3	–	$t7$	$t8$
4	6	4	Elsa	$t9$	$t10$
5	6	5	Anna	$t10$	$t10$

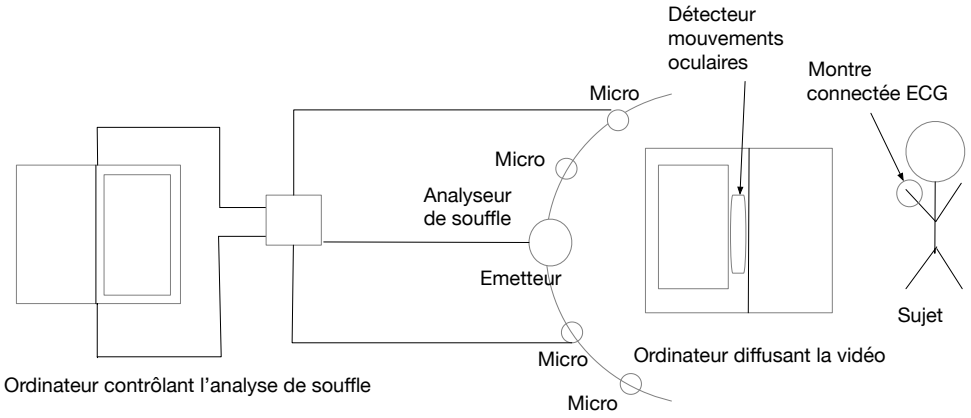

Figure 1: Dispositif de récolte des signaux physiologiques

Les données collectées par le détecteur de mouvements oculaires sont la taille des deux pupilles et la position de la fixation de chaque oeil. A partir de ces données, nous avons calculé les valeurs des descripteurs qui sont utilisés pour la reconnaissance des émotions : taille des pupilles [13], position de l'oeil, nombre de clignements de yeux et leur durée moyenne[14], les points de fixation [15], la durée des fixations[16], nombre et durée moyenne des saccades: cela correspond au mouvement rapide des yeux entre deux positions [17]. En plus de ces descripteurs, est ajouté l'angle de rotation entre le point de départ et les deux saccades successives afin de prendre en compte le changement de direction du regard. Le capteur ECG de la montre enregistre les variations dans le potentiel électrique de la membrane myocarde, un signal en volts en fonction du temps. Nous extrayons une représentation issue d'un modèle pré-entrainé du signal filtré et normalisé en utilisant une architecture neuronale et un modèle pré-entraîné [18].

Les données actuelles d'EMOLIS Dataset correspondent actuellement à seulement quelques personnes mais l'idée n'est pas d'obtenir un modèle physiologique générique mais de montrer la faisabilité de l'approche. Nous avons les données de sept spectateurs, y compris les auteurs.

Table 5
Fréquence de chaque émotion et l'émotion la plus représentative exprimées par un annotateur de la scène 1 de Mulan.l'annotateur a ressenti en premier la joie (j) 3 fois, puis le dégoût (d), puis à nouveau la joie trois fois et enfin la tristesse (t). L'émotion la plus représentative est la joie. La colère (c), la peur (p), la surprise (s) and autre (a) ne sont pas ressenties. (n) correspond à "pas d'émotions ressenties".

		Fréquence des émotions							émotion majoritaire							
film	scène	c	d	j	p	s	t	a	c	d	j	p	s	t	a	n	séquence
Mulan	1	0	1	6	0	0	1	0	0	0	1	0	0	0	0	0	$j\ j\ j\ d\ j\ j\ j\ sa$

Table 6
Distribution des émotions majoritaires dans les scènes.

émotion majoritaire	j	c	p	t	d	s	a
# des scènes	19	7	17	16	5	4	0

Table 7
Score Kappa de Fleiss entre les annotateurs (les émotions les plus représentatives).

Fusion	j	c	p	t	d	s	a
Kappa	0.4	0.4	0.5	0.6	0.3	0.1	0

3.3. Annotation de l'émotion ressentie

Après avoir enregistré les signaux, le spectateur visionne à nouveau l'ensemble des scènes pour annoter les émotions qu'il estime ressentir. Pour chaque scène, chaque annotateur doit identifier la chronologie de ses émotions. Puis, il doit déterminer la ou les émotions les plus représentative(s) pour lui ou indiquer "autre" (a) ou "aucune"(n). La durée d'annotation est d'environ 5h20mn pour 2h 8mn de scènes. Les annotations des spectateurs sont dans des fichiers CSV. Les lignes contiennent le numéro de la scène, la fréquence de chaque émotion ressentie, l'émotion la plus représentative et la séquence d'apparition des émotions (voir table 5). Les annotateurs sont issus de milieux et de niveaux d'études variés de 12 à 64 ans et sont francophones. Deux d'entre eux sont autistes Asperger, âgés de 23 et 27 ans.

Sur chaque scène, nous appliquons un vote majoritaire sur les émotions les plus représentatives de tous les annotateurs afin d'avoir une vérité terrain. En effet, les émotions étant subjectives, des différences sont attendues et normales. Mais si la majorité des annotateurs dit avoir ressenti la même émotion, nous considérons ce ressenti comme vérité terrain.

Pour deux scènes, le vote majoritaire sur les émotions les plus représentatives des annotateurs n'a pas permis d'annoter comme vérité terrain une seule émotion. Nous avons alors retenu les deux émotions ayant le même nombre de votes. La Table 6 montre la distribution des émotions les plus représentatives dans les scènes. On remarque que les émotions sont bien distribuées parmi joie (j), peur (p) et tristesse (t), mais pas parmi colère (c), surprise (s) et dégoût (d). Nous obtenons approximativement le même nombre de scènes attribuées à chaque émotion.

Les Kappa de Fleiss [19] entre annotateurs sont présentés pour chaque émotion dans la table 7. Ce Kappa mesure l'accord entre plusieurs annotateurs, contrairement au Kappa de Cohen

[20] qui mesure la concordance des observateurs deux à deux. Le score Kappa est moyen pour joie (0,4), colère et peur (0,5), faible pour dégoût (0,3) et surprise (0,1) et important pour tristesse (0,6). En comparaison, le score de Kappa du jeu de données de MELD [11] est de 0,43. On peut en déduire que l'impact des facteurs socioculturels et personnels est important [21]. Par exemple, parmi nos annotateurs, quatre sont membres de la même famille et le score de Kappa de Cohen, calculé deux à deux, est important (0,61 à 0,78) pour la peur et moyen à important (0,53 à 0,74) pour la tristesse. alors que pour ceux qui ne font pas partie d'une même famille, les scores varient fortement de léger à important (0,2 à 0,58) pour la peur et (0,35 à 0,65) pour la tristesse.

La surprise peut être confondue par les annotateurs avec la joie et la peur. En effet„ la surprise peut être considérée comme un mélange de deux émotions puisque si la surprise est agréable, elle peut être associée à la joie et si la surprise est désagréable, elle peut être associée à la peur. Nous pensons que la surprise est associée à un moment court de déstabilisation, de peur. Si l'annotateur est rassuré par la raison, la joie remplace ensuite la surprise. Si la surprise est très forte et que la raison confirme que l'état de déstabilisation et de peur est normal, la peur devient plus intense et remplace donc la surprise. Le dégoût peut être également confondu avec la colère et la tristesse. Si le dégoût n'est pas déclenché par les cinq sens mais par le comportement d'un personnage, il peut être associé à la tristesse et à la colère car l'annotateur est déçu du personnage auquel il a été attaché auparavant. L'annotateur peut aussi ressentir de l'aversion si le comportement du personnage est contraire à ses valeurs, ce qui l'amène à le rejeter.

Par ailleurs, la variabilité entre les annotateurs peut être due aux difficultés de perception des émotions des personnes autistes Asperger et/ou la difficulté à émettre une émotion représentative de la scène ou encore aux écarts d'âge des spectateurs. La nature subjective de cette tâche peut rendre vaine la volonté d'établir un consensus. Au mieux, cela nécessiterait d'avoir un panel d'annotateurs varié et important ou encore de les catégoriser selon des critères à déterminer.

4. EMOLIS App

Dans cette section, nous présentons EMOLIS App (le code source sera distribué librement). Elle permet d'estimer en temps réel les émotions émises par une vidéo d'EMOLIS Dataset choisie par l'utilisateur et suggère des scènes émotionnellement similaires, selon le contenu et les signaux physiologiques (voir figure 2). A partir d'une vidéo d'EMOLIS Dataset, les émotions sont prédites en temps réel. Nous récupérons également avant la classification les représentations transformeurs de l'image, de l'audio et du texte pour les exploiter lors de la recommandation conjointement avec les descripteurs des signaux physiologiques (voir figure 3).

Les émotions émises par la vidéo sont prédites à partir d'une architecture neuronale dérivée de Siriwardhana et al. [22]. Elle utilise des représentations pré-entraînées pour les modalités audio (Wav2vec), texte (Roberta) et image (Fabnet Video). Pour obtenir des représentations pour l'audio et l'image (jetons CLS), des transformeurs unimodaux sont entraînés et utilisés pour identifier les émotions à partir des représentations texte, audio et image. Parce que l'audio des films contient des bruitages et de la musique en plus de la voix, nous isolons cette dernière grâce à Spleeter [23]. Afin de croiser les modalités, des transformeurs bimodaux sont appliqués afin d'obtenir six représentations bimodales. Après avoir appliqué un produit d'Hadamar à chaque paire, nous concaténons les représentations. Le modèle est ensuite entraîné sur celles-ci

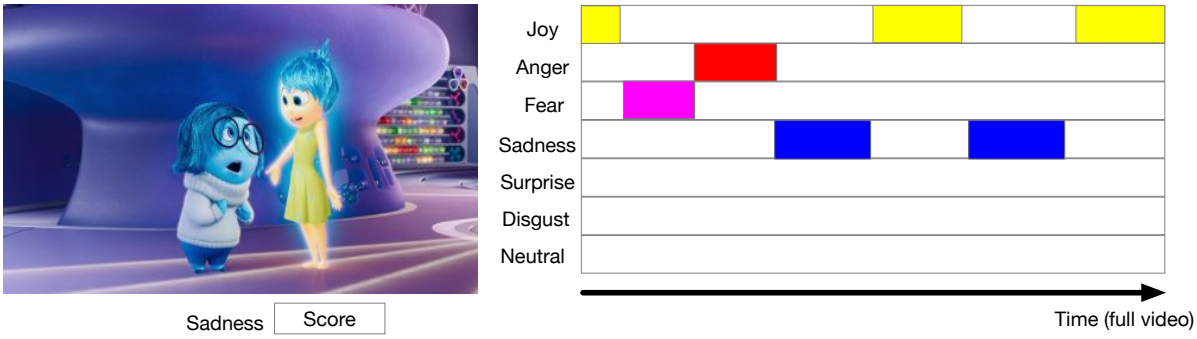

Figure 2: Vue d'EMOLIS App

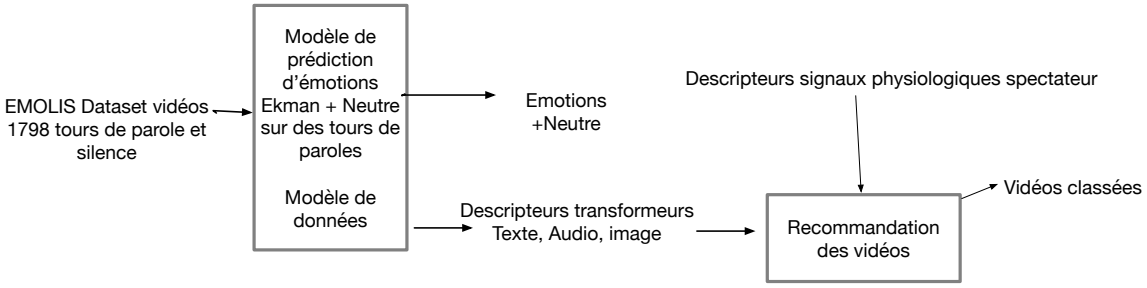

Figure 3: Architecture d'EMOLIS App

avec des couches linéaires pour la prédiction de chaque émotion. Nous obtenons une F-mesure de 62% pour la prédiction d'émotions du jeu de données de test sur MELD, ce qui est comparable aux résultats mentionnés par Siriwardhana et al. [22]. En calculant une distance "émotionnelle" entre deux vidéos, nous utilisons comme Rouabhia and Tebbikh [24], la distance dérivée de la norme de Frobénius sur les représentation matricielles associées (quatre matrices) à chaque type de signaux, aussi bien qu'aux représentations des contenus. La distance émotionnelle entre deux scènes est vue comme la somme des distances. EMOLIS App peut donc suggérer des vidéos à partir des représentations les plus proches. L'utilisateur peut évaluer la suggestion en notant la similarité entre la vidéo qu'il a choisie et la vidéo suggérée (de 1 à 4). Nous pouvons également évaluer la suggestion à partir des annotations du spectateur.

5. Conclusion

Nous avons d'abord construit EMOLIS Dataset, un nouveau jeu de données multimodales issues de dessins animés populaires, annotées avec des émotions et associés à des signaux physiologiques. Par ailleurs, nous avons considéré les signaux physiologiques et le contenu des vidéos pour suggérer de manière personnalisée, à partir de représentations établies par des architectures et modèles neuronaux, des vidéos émotionnellement similaires. Nous utiliserons EMOLIS App dans le contexte des troubles de la théorie de l'esprit et pour l'entraînement émotionnel des jeunes personnes atteintes du trouble du spectre autistique.

References

[1] P. Ekman, An argument for basic emotions, Cognition & emotion 6 (1992) 169–200.

[2] G. Atherton, L. Cross, Seeing more than human: autism and anthropomorphic theory of mind, Frontiers in psychology 9 (2018) 528.

[3] C. Tardif, L. Latzko, T. Arciszewski, B. Gepner, Reducing information's speed improves verbal cognition and behavior in autism: A 2-cases report, Pediatrics 139 (2017).

[4] S. Farashi, S. Bashirian, E. Jenabi, K. Razjouyan, Effectiveness of virtual reality and computerized training programs for enhancing emotion recognition in people with autism spectrum disorder: a systematic review and meta-analysis, International Journal of Developmental Disabilities (2022) 1–17.

[5] N. Jain, V. Gupta, S. Shubham, A. Madan, A. Chaudhary, K. Santosh, Understanding cartoon emotion using integrated deep neural network on large dataset, Neural Computing and Applications (2021) 1–21.

[6] C. Busso, M. Bulut, C.-C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J. N. Chang, S. Lee, S. S. Narayanan, Iemocap: Interactive emotional dyadic motion capture database, Language resources and evaluation 42 (2008) 335–359.

[7] J. A. Russell, A circumplex model of affect., Journal of personality and social psychology 39 (1980) 1161.

[8] Y. Baveye, E. Dellandrea, C. Chamaret, L. Chen, Liris-accede: A video database for affective content analysis, IEEE Transactions on Affective Computing 6 (2015) 43–55.

[9] M. Soleymani, J. Lichtenauer, T. Pun, M. Pantic, A multimodal database for affect recognition and implicit tagging, IEEE transactions on affective computing 3 (2011) 42–55.

[10] L. Santamaria-Granados, M. Munoz-Organero, G. Ramirez-Gonzalez, E. Abdulhay, N. Arunkumar, Using deep convolutional neural network for emotion detection on a physiological signals dataset (amigos), IEEE Access 7 (2018) 57–67.

[11] S. Poria, D. Hazarika, N. Majumder, G. Naik, E. Cambria, R. Mihalcea, Meld: A multimodal multi-party dataset for emotion recognition in conversations, arXiv preprint arXiv:1810.02508 (2018).

[12] P. Arlotto, M. Grimaldi, R. Naeck, J.-M. Ginoux, An ultrasonic contactless sensor for breathing monitoring, Sensors 14 (2014) 15371–15386.

[13] L. Moharana, N. Das, Analysis of pupil dilation on different emotional states by using computer vision algorithms, in: 2021 1st Odisha International Conference on Electrical Power Engineering, Communication and Computing Technology (ODICON), IEEE, 2021, pp. 1–6.

[14] P. R. K. Babu, U. Lahiri, Classification approach for understanding implications of emotions using eye-gaze, Journal of Ambient Intelligence and Humanized Computing 11 (2020) 2701–2713.

[15] M. Deng, X. Gu, Information acquisition, emotion experience and behaviour intention during online shopping: an eye-tracking study, Behaviour & Information Technology (2020) 1–11.

[16] E. S. Martínez-Velázquez, A. L. Ahuatzin González, Y. Chamorro, H. Sequeira, The influence of empathy trait and gender on empathic responses. a study with dynamic emotional stimulus and eye movement recordings, Frontiers in psychology 11 (2020) 23.

[17] L. J. Schmidt, A. V. Belopolsky, J. Theeuwes, The presence of threat affects saccade trajectories, Visual Cognition 20 (2012) 284–299.
[18] P. Sarkar, A. Etemad, Self-supervised ecg representation learning for emotion recognition, IEEE Transactions on Affective Computing 13 (2020) 1541–1554.
[19] J. L. Fleiss, Measuring nominal scale agreement among many raters., Psychological bulletin 76 (1971) 378.
[20] J. Cohen, A coefficient of agreement for nominal scales, Educational and psychological measurement 20 (1960) 37–46.
[21] H. W. Krohne, Individual differences in emotional reactions and coping. (2003).
[22] S. Siriwardhana, T. Kaluarachchi, M. Billinghurst, S. Nanayakkara, Multimodal emotion recognition with transformer-based self supervised feature fusion, IEEE Access 8 (2020) 176274–176285.
[23] R. Hennequin, A. Khlif, F. Voituret, M. Moussallam, Spleeter: a fast and efficient music source separation tool with pre-trained models, Journal of Open Source Software 5 (2020) 2154. Deezer Research.
[24] C. Rouabhia, H. Tebbikh, Mesure de similarité pondérée dans l'espace 2d: Application à la reconnaissance de visages., in: CORIA, 2010, pp. 373–385.