Analyse multimodale de scène : vers une intégration des données contextuelles ?

Ibrahim Mohamed Serouis

Analyse multimodale de scène : vers une intégration des données contextuelles?

Ibrahim MOHAMED SEROUIS

Université Paul Sabatier, Laboratoire IRIT
118 Route de Narbonne
31062 Toulouse

ibrahim.mohamed-serouis@irit.fr

RÉSUMÉ : *Dans divers domaines comme la communication, le cinéma et les interactions humaines, l'avènement de l'apprentissage multimodal ouvre de nouvelles perspectives dans l'analyse des interactions, des scènes et des publicités. Cependant, peu d'études explorent les modalités autres que l'image et l'audio, négligeant des informations contextuelles cruciales. Cette étude propose un modèle de données pour l'analyse de scènes centrées sur l'humain et une méthodologie pour automatiser l'extraction de ces données à partir de vidéos, ainsi qu'une méthode pour intégrer les données contextuelles dans l'analyse multimodale des scènes.*

MOTS-CLÉS : *Analyse de scène, Apprentissage multimodal, Modélisation de données, Extraction de données multimédias.*

ENCADREMENT : *Florence SÈDES (PR), Lucile SASSATELLI (PR)*

1. Introduction et problématique

Dans les domaines de la communication, du cinéma, ou lors d'interactions humaines, les canaux de communication peuvent être à la fois visuels, textuels, auditifs et contextuels. L'émergence croissante de l'apprentissage multimodal a ouvert de nouvelles perspectives, notamment en ce qui concerne l'analyse d'interactions, de scènes ou de publicités, avec des résultats de plus en plus prometteurs (Schauerte *et al.*, 2011) (Xu *et al.*, 2016) (Gasparini *et al.*, 2018) (Kukleva *et al.*, 2020).

Cependant, très peu d'études à l'instar des travaux de (Gasparini *et al.*, 2018) et (Vicol *et al.*, 2018) exploitent des modalités autres qu'une image associée à sa

description et/ou un audio, négligeant ainsi les informations contextuelles telles que le contexte de l'interaction ou les relations entre les personnages à l'écran. Or, ne pas tenir compte du contexte pourrait entraîner une perte d'informations cruciales lors de l'analyse d'une interaction, le point de vue pouvant être influencé par exemple par le caractère formel ou non de la situation, ou encore de la relation entre les acteurs de l'interaction. De même, rares sont les études exploitant l'aspect relationnel entre les différentes informations en entrée, et encore plus celles mettant un accent sur l'extraction et/ou la représentation de ces données, à l'instar de (Al-Jarrah *et al.*, 2015), (Panta *et al.*, 2018), et plus récemment (Qodseya, 2020).

Cette étude a pour objectif d'ouvrir de nouvelles perspectives pour une compréhension plus approfondie des situations humaines par les systèmes automatisés. Pour ce faire, nous proposons, d'une part, un modèle de données permettant la représentation des données pour un problème d'analyse de scène centrée sur l'humain, ainsi qu'une méthodologie permettant d'automatiser l'extraction de ces données à partir d'une vidéo. D'autre part, nous proposerons une méthodologie interprétable permettant d'intégrer les données contextuelles à l'analyse multimodale de scène.

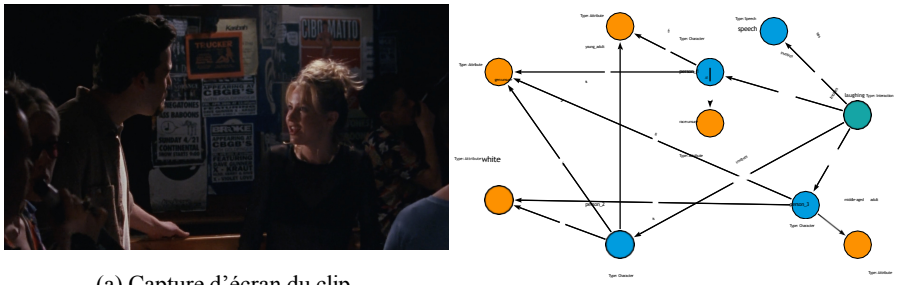
2. Actions réalisées et limites

2.1. *Modèle de données et extraction des données*

Dans le cadre de cette étude, une première itération du module AMDER (acronyme de **A**dvancing **M**ultimedia **D**ata **E**xtraction and **R**epresentation) a été développée. Comme illustré dans les figures 1a et 1b, ce module prend en entrée une vidéo en et génère en sortie un graphe de scène. Ce graphe comprend les interactions vidéo, les coordonnées de détection des personnages, ainsi que leurs attributs physiques et les émotions exprimées au cours de la scène. Ce module peut être utilisé comme un outil de pré-annotation, auquel on pourrait appliquer un algorithme d'analyse de relation entre les personnages, tel que celui mentionné dans la section 2.2, afin d'enrichir les données contextuelles. La sortie du module est une représentation qui s'inscrit dans une première tentative de modélisation des données relatives aux scènes. Cette modélisation inclut les personnages, leurs attributs (tels que le sexe, la race, les attributs particuliers), les coordonnées de détection, les émotions exprimées pendant la scène, l'ensemble des interactions réalisées dans une scène, ainsi que le discours tenu lors de l'interaction.

2.2. *Modèle d'apprentissage*

Les approches basées sur les Graph Neural Networks (GNNs) sont de plus en plus populaires pour les problèmes de classification sur des données multimodales, bien que certaines réticences aient été exprimées dans la littérature (Ektefaie *et al.*, 2023). Les GNNs permettent d'exploiter l'aspect relationnel entre les données et de traiter des données de tailles variables, ce qui en fit une piste intéressante à explorer pour notre



(a) Capture d'écran du clip

(b) Sortie du module : graphe de scène

Figure 1: Exemple d'exécution du module AMDER sur un extrait de *Chasing Amy*.

problème. Nous avons donc développé une méthodologie en trois étapes pour tirer parti des données contextuelles. La première étape consiste en l'encodage des données d'entrée, via un modèle d'*embedding*. La deuxième étape consiste en l'apprentissage de la représentation des noeuds du graphe. La dernière étape consiste en la classification des noeuds d'intérêt (scène, interaction), et la génération de statistiques sur la contribution des noeuds au résultat final. Cette méthodologie a été évaluée sur différentes tâches d'analyse de scène, telles que la détection d'objectification, la classification d'interactions et la détection de relations entre les personnages. Les résultats obtenus sont prometteurs, dépassant même ceux de (Kukleva *et al.*, 2020) sur la classification d'interactions et de relations entre les personnages, sur un même jeu de données. Pour résoudre le problème de boîte noire (Hussain, 2019) associé à ce type d'algorithme, nous avons également proposé une approche permettant d'obtenir l'influence de chaque élément de notre graphe. Néanmoins, cette dernière reste perfectible notamment concernant les choix d'architecture et d'opérations de traitement des données.

3. Conclusions et perspectives

Cette étude a pour objectif d'ouvrir de nouvelles perspectives pour une compréhension plus approfondie des situations humaines par les systèmes automatisés.

Dans un premier temps, nous présentons un module d'extraction de données vidéo, qui peut servir de pré-annotation pour la création de jeux de données contenant des scènes. Les premiers résultats étant globalement satisfaisants, nous envisageons d'intégrer des techniques de Human Parsing, telles que celle proposée par (Liang *et al.*, 2018), afin d'obtenir des détails plus fins sur les tenues vestimentaires des personnages à l'écran. Cela permettrait par exemple de détecter la nudité, ou une tenue inappropriée dans un contexte sérieux. Nous prévoyons également d'ajouter des techniques de réduction de bruit pour améliorer l'extraction du discours dans les vidéos, ainsi que l'utilisation de grands modèles de langage pour obtenir une description de la scène basée sur les éléments extraits.

Enfin, nous proposons une méthodologie interprétable de raisonnement sur les scènes, qui peut intégrer des données contextuelles (comme les relations entre les personnages, le lieu de la scène) en plus des entrées traditionnelles (images, transcription du discours). Bien que cette méthodologie soit plus performante que certaines méthodes de référence, telles que celle proposée par (Kukleva *et al.*, 2020), pour la classification d'interactions, elle pourrait être améliorée en intégrant des connaissances métiers, en particulier pour le problème de détection d'objectification. Nous envisageons donc d'introduire une approche neuro-symbolique qui bénéficierait des retours d'experts pour la partie symbolique, et des sorties de notre modèle comme connaissances préalables.

4. Bibliographie

Al-Jarrah O. Y., Yoo P. D., Muhaidat S., Karagiannidis G. K., Taha K., "Efficient machine learning for big data: A review", *Big Data Research*, vol. 2, n° 3, p. 87–93, 2015. Publisher: Elsevier.

Ektefaie Y., Dasoulas G., Noori A., Farhat M., Zitnik M., "Multimodal learning with graphs", *Nature machine intelligence*, vol. 5, n° 4, p. 340–350, avril, 2023.

Gasparini F., Erba I., Fersini E., Corchs S., "Multimodal Classification of Sexist Advertisements:", *Proceedings of the 15th International Joint Conference on e-Business and Telecommunications*, SCITEPRESS - Science and Technology Publications, Porto, Portugal, p. 399–406, 2018.

Hussain J.*, "Deep Learning Black Box Problem"*, Master's thesis, Uppsala University, Department of Informatics and Media, 2019. Backup Publisher: Uppsala University, Department of Informatics and Media.

Kukleva A., Tapaswi M., Laptev I., "Learning Interactions and Relationships between Movie Characters", 2020. arXiv:2003.13158 [cs].

Liang X., Gong K., Shen X., Lin L., "Look into Person: Joint Body Parsing & Pose Estimation Network and a New Benchmark", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2018. Publisher: IEEE.

Panta F. J., Roman-Jimenez G., S'edes F., "Modeling metadata of CCTV systems and Indoor Location Sensors for automatic filtering of relevant video content", *2018 12th International Conference on Research Challenges in Information Science (RCIS)*, IEEE, p. 1–9, 2018.

Qodseya M., Managing heterogeneous cues in social contexts. A holistic approach for social interactions analysis, PhD thesis, Universite Toulouse 3 Paul Sabatier, 2020.

Schauerte B., Kühn B., Kroschel K., Stiefelhagen R., "Multimodal saliency-based attention for object-based scene analysis", *2011 IEEE/RSJ International Conference on Intelligent Robots and Systems*, p. 1173–1179, 2011.

Vicol P., Tapaswi M., Castrejon L., Fidler S., "MovieGraphs: Towards Understanding Human-Centric Situations from Videos", *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

Xu P., Davoine F., Bordes J.-B., Zhao H., Denœux T., "Multimodal information fusion for urban scene understanding", *Machine Vision and Applications*, vol. 27, n° 3, p. 331–349, avril, 2016.