Vers une approche inclusive de description multimédia pour le patrimoine

LILIA DJOUSSOUF[1], KATERINE ROMEO[1], ABDERRAHIM EL MOATAZ BILLAH[2]

[1]LITIS, Avenue de l'Université, 76800 Saint-Etienne-du-Rouvray, France
[2] GREYC, 6 Boulevard Maréchal Juin, 14000 Caen, France

Résumé – Rendre accessible le patrimoine culturel pour créer un espace public inclusif est un enjeu majeur pour les institutions telles que les musées. Les avancées techniques en vision par ordinateur, et apprentissage automatique montrent actuellement des évolutions possibles. Dans cette optique, cet article cherche à établir une chaîne de traitements possibles se basant sur la perception visuelle de personnes à vue normative pour l'extraction d'informations sémantiquement importantes. L'application porte sur des images issues de la Tapisserie de Bayeux. Ces zones saillantes sémantiquement importantes peuvent être adaptées sur un dispositif haptique, nommée F2T (force feedback tablet) à travers l'extraction de contours multi-échelles. Enfin, une piste sur l'utilisation d'un modèle récent de diffusion latente permettant de générer des effets sonores à partir d'une description est évoquée.

Abstract – Making cultural heritage accessible to create an inclusive public space is a major issue for institutions such as museums. Technical advances in computer vision and machine learning show the possible evolution. In this perspective, this article seeks to establish a possible pipeline based on the visual perception of sighted people for the extraction of semantically important information. The method is applied on the images from the Bayeux Tapestry. These semantically important salient areas could be adapted on a haptic device, named F2T (force feedback tablet), through the extraction of multi-scale contours. Finally, a possible use of a recent latent diffusion model allowing to generate sound effects from a description is mentioned.

1 Contexte et motivations

Le patrimoine culturel matériel, tel que la Tapisserie de Bayeux[1], est difficilement accessible aux personnes aveugles et partiellement aveugles. Parmi les personnes qui voient, il est estimé qu'un visiteur passe en moyenne entre 17 et 21 secondes devant une œuvre d'art [1, 2]. Rendre accessible le patrimoine en intégrant des représentations multimédias serait une solution possible pour intéresser les visiteurs aveugles, partiellement aveugles et qui voient.

La Tapisserie de Bayeux est une broderie s'étalant sur environ 70m de longueur avec 50cm de hauteur. Cette œuvre datant du XIe siècle est inscrite au patrimoine mondial selon l'UNESCO. Elle est complexe et comporte un grand nombre d'éléments sémantiquement importants.

Les avancées en traitement de l'information pourraient permettre de répondre aux besoins des visiteurs de musée. Ceci en passant par des algorithmes d'apprentissage automatique autour de la vision par ordinateur, du traitement de langage naturel et de l'audio.

La figure 1 correspond à une vue globale pour une chaîne de traitement à réaliser pour adapter des images pouvant être explorée avec des retours audios et haptiques via l'interface de la F2T.

Cet article propose d'étudier plusieurs étapes de cette chaîne de traitement permettant d'adapter des images issues de la Tapisserie de Bayeux en représentations multimédias en y exposant les méthodes et les limites rencontrées. Dans ce cadre, la section 2 donne un aperçu de résultats pour l'extraction de zones sémantiquement importantes par suivi oculaire pour des images issues de la Tapisserie de Bayeux, avec les limites du protocole suivi. La section 3 présente des applications de segmentation d'images dans le but d'extraire des contours utiles pour la création de représentations supportées par un dispositif haptique (F2T). Enfin, la section 4 conclut l'article avec des perspectives.

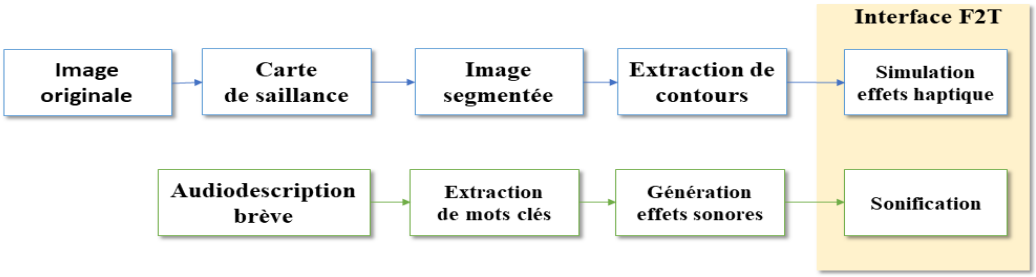

Figure 1. Vue globale d'une chaîne de traitement pour la perception d'images via l'interface de la F2T

[1] Site web: https://www.bayeuxmuseum.com/la-tapisserie-de-bayeux/, dernière consultation : 07.04.2023

(a) (b) (c)

Figure 2. (a) Zones d'intérêts définis manuellement pour la scène 23 de la Tapisserie de Bayeux, le « Serment de Harold ». (b) Cartes de chaleur associées à cette scène par le participant ET-1 avec un stimulus audio correspondant à la description détaillée, (c) participant ET-2 sans stimulus audio.

2 Extraction d'informations sémantiquement importantes

Dans le cadre d'images complexes telles que la Tapisserie de Bayeux, une grande quantité d'éléments notables sont pris en compte. L'extraction de zones saillantes selon les caractéristiques seules de la couleur et la luminosité ne suffit pas, car il existe un risque de perte d'éléments sémantiquement importants. Cela revient au problème de la vision par ordinateur qui cherche à se rapprocher de la manière qu'un être humain visualise une image.

Yarbus [3] a pu démontrer que selon une tâche à accomplir le regard peut être guidé pour chercher les informations jugées intéressantes.

Plus récemment, Reinholt *et al.* [4] proposent un système d'annotation, EyeDescribe. Ce système traite les données issues de suivi oculaire et de paroles dans le but d'annoter des images (de peinture, carte etc..) pour des personnes aveugles et partiellement aveugles. A partir des données issues de suivi oculaire, il est possible de détecter des zones d'intérêt dans une image, synchronisées avec les paroles enregistrées d'experts conversant sur ces images, les régions sont ainsi décrites. Ces images annotées sont alors rendues accessibles via un écran tactile et un support en bois pour localiser les bords de l'écran et une voix synthétisée pour lire les descriptions. Une des limites rencontrées lors de l'utilisation par le public visé, est la difficulté liée à localiser spatialement les éléments.

Il est possible de retrouver dans la littérature des jeux de données tels que Localized Narratives [5]. Ils ont la particularité de rassembler des descriptions textuelles, audio, et la position des éléments décrits par des germes fournis par l'annotateur avec une souris d'ordinateur. Ce jeu de données a pour avantage de pouvoir être utilisé pour différentes tâches multimodales (e.g., image vers texte, texte vers image, texte vers audio, etc.). Les images constituant ce jeu de données proviennent de quelques images issues des bases de données suivantes : Open Images, COCO, Flickr30k et ADE20k. Ces images représentent majoritairement des scènes naturelles. Il existe peu de bases de données composées d'images de peintures assez bien annotées. C'est pourquoi, nous souhaitons élaborer un protocole permettant de construire un jeu de données d'images du même type que la Tapisserie de Bayeux, où les informations sur l'attention et la sémantique seraient disponibles.

Une étude pilote a été réalisée pour observer le comportement du regard face à une image complexe telle que la Tapisserie de Bayeux, avec et sans stimuli audio, sous la forme d'audiodescription. Ce qui est recherché ici est le degré d'attention attribué à des zones d'intérêts renseignées par des experts. La figure 2 présente un exemple de résultats obtenus par l'utilisation d'un eye-tracker pour une scène de la Tapisserie de Bayeux. La figure 2.a correspond à une image dans lesquelles les zones d'intérêts équivalentes aux zones sémantiquement importantes de l'image sont numérotées de 1 à 4 Les zones d'intérêts définies manuellement sont {Guillaume ; Harold ; reliquaire 1 ; reliquaire 2}.

Pour cette étude, l'eye-tracker basé sur écran, Tobii Pro Spark, a permis d'obtenir des cartes de chaleur de deux participants identifiés par ET-1 et ET-2.

L'audiodescription détaillée de la scène dure 3min50s, elle a été coécrite avec des personnes partiellement aveugles et avec vision normative détaillée (ICAD, Inclusive Co-created Audio Description). Le participant ET-1 a visualisé l'image en étant guidé par cette audiodescription inclusive. Le participant ET-2, quant à lui, a visualisé l'image sans stimuli audio pendant la même durée.

L'objectif était d'observer l'influence de la description orale liée à l'image sur l'attention visuelle.

Pour pouvoir comparer les résultats, les deux participants ont dû visualiser l'image pendant le même intervalle de temps.

Le tableau 1 présente le nombre de fixations par zones d'intérêts. A titre comparatif, il est possible de remarquer que le nombre de fixations est de 124 et 94 fixations pour le participant ET-1, et 84 et 22 fixations pour le participant ET-2, respectivement pour le reliquaire 1 (zone 3) et le reliquaire 2 (zone 4). Sans guide audio, il semblerait que l'attention portée aux zones d'intérêts soit plus dispersée.

Tableau 1. Nombre de fixations relevé par zones d'intérêt pour les participants ET-1 et ET-2.

	Nombre de fixations par zones d'intérêts				
Zones d'intérêts / Participants	Guillaume (1)	Harold (2)	Relique (3)	Relique (4)	Total
ET-1	115 (26%)	110 (25%)	124 (28%)	94 (21%)	443 (100%)
ET-2	78 (24%)	142 (43%)	84 (26%)	22 (7%)	326 (100%)

Les descriptions orales utilisées ici sont des audiodescriptions détaillées, elles sont donc exhaustives

en information. Ce qui limite le nombre d'images et de descriptions orales pouvant être testées avec des participants, et ainsi introduire une fatigue trop importante. De plus, les zones d'intérêt indiquées pouvant être trop proches, voire imbriquées, il est plus compliqué de les délimiter via le logiciel de l'eye tracker. Pour résoudre ce premier problème, une voie envisageable est d'utiliser des descriptions plus concises, de l'ordre de 1 à 2 phrases maximum pour décrire une image.

Dans cette section, nous avons considéré des améliorations pour un protocole pour construire un jeu de données utilisant des informations sur l'attention et la sémantique. La section 3 présentera des applications possibles pour obtenir des représentations multimédia pour des images de la Tapisserie de Bayeux.

3 Application : représentations multimédia

3.1 F2T – dispositif haptique support de représentations multimédias

A partir de ces éléments saillants et sémantiquement importants, l'objectif est de les représenter sur un support spécifique supportant différentes modalités (audio et haptique) : la F2T (Force Feedback Tablet) [6]. Cette tablette à retour d'effort (Fig. 3) permet l'exploration libre et guidée avec un joystick à partir d'une image virtuelle. Le logiciel de la F2T permet de créer une image virtuelle dont les propriétés peuvent modifier la vitesse du joystick, mais aussi, selon sa localisation, émettre des audiodescriptions ou des effets sonores.

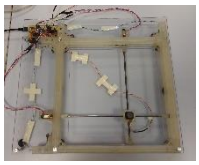

Figure 3. F2T, la tablette à retour de force[2].

Dans le cadre d'éléments figuratifs, il est nécessaire, dans un premier temps, de segmenter l'image. Les contours participeraient à la perception des formes en attribuant des effets haptiques. Les ICAD ou des effets sonores seraient alors mis à disposition pour faciliter l'intégration cognitive des informations sémantiques liées à l'œuvre.

3.2 Méthodes de segmentation d'images pour l'adaptation à des représentations tactiles

Way et Barner [7] ont été les précurseurs pour proposer une automatisation de la simplification d'images pour les rendre accessibles aux personnes aveugles. Pour cela, ils proposent d'utiliser une chaîne de traitement utilisant des techniques de traitement d'images, et ainsi obtenir les contours des éléments présents dans l'image. D'après les auteurs, les contours simplifiés pourraient donc être utilisés pour être appliqués pour des supports tactiles statiques (e.g., thermogonflés) ou dynamique (e.g., dispositif braille).

Abdusalomov *et al.* [8] proposent notamment d'automatiser le processus de création de représentation tactile d'images en se basant sur la saillance des régions dans l'image pour l'extraction et en adaptant automatiquement les paramètres de l'algorithme de Grabcut [9]. Néanmoins, comme précisé par l'auteur, il est difficile d'extraire les régions saillantes dans l'image lorsque les régions appartenant à l'arrière-plan et le premier plan sont trop similaires.

La figure 4 présente des techniques de segmentation d'images avec extraction des contours appliquées à la Tapisserie de Bayeux. Les contours extraits serviraient notamment à être présentés sur des supports tactiles ou dynamique pour ainsi permettre leur exploration haptique.

Deux méthodes non supervisées ont été appliquées sur des zones d'intérêts sur des images issues de la Tapisserie de Bayeux. L'objectif étant d'obtenir des contours simplifiés pour ainsi limiter la charge cognitive lors du suivi des contours. La première approche se base sur la segmentation de l'image par un algorithme de meanshift en appliquant au préalable un filtre moyenneur adaptatif. Les points de frontières des

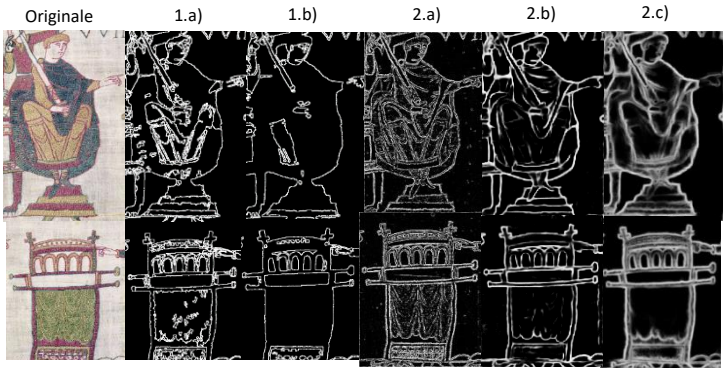

Figure 4. Comparaison de méthodes de segmentation d'images pour l'extraction de contours (1) segmentation d'image par meanshift avec filtre gradient 1.a) $\sigma = 1$ et 1.b) $\sigma = 5$; (2) Détection de contours par Holistically Nested Edge Detection (HED) : 2.a) contours intermédiaire de la couche 2 ; 2.c) contours intermédiaire de la couche 4 ; 2.c) fusion des cartes de contours

[2] Site web: https://gaysimon.github.io/postdoc/F2T2.html , dernière consultation : 14.04.2023

régions obtenues ont été localisés et puis extraits (cf. figure 4.1.a et 4.1.b). Il est difficile avec les méthodes classiques de traitement d'images de contrôler le niveau de détails à extraire ; une étape de post traitement est encore nécessaire pour obtenir uniquement les contours souhaités. La seconde approche se base sur un modèle d'apprentissage profond permettant d'extraire les contours multi-échelles : HED (Holistically Nested-Edge Detection) [10]. Il a été entraîné sur le jeu de données BSDS500. C'est un modèle d'apprentissage profond se basant sur VGG16 et fusionnant les images résultantes de différentes couches (fig. 4.2.a et 4.2.b) pour obtenir la fusion des cartes de contours (fig. 4.2.c). En obtenant des images présentant des contours à plusieurs échelles pouvant ensuite être retranscrit pour le dispositif haptique, les utilisateurs pourraient sélectionner un niveau de détail précis.

En comparant qualitativement les images résultantes pour ces deux méthodes, il est possible d'observer que l'approche avec HED (fig. 4.2.c) présente des contours plus saillants en gardant les contours accentués à plusieurs niveaux de gris. Cette méthode a l'avantage de ne pas avoir à ajuster le modèle, comparé à la méthode se basant sur la segmentation d'images par meanshift qui présente une perte de contours lors de l'ajustement de la taille d'un masque de convolution d'un filtre gradient.

Cette partie a permis de comparer les méthodes de segmentation d'image pour l'extraction de contours simplifiés multi-échelles.

4 Conclusion et perspectives

Dans cet article il a été possible de mettre en avant diverses techniques pouvant servir à l'élaboration de représentations multimédia pouvant être supportées sur une tablette à retour de force, F2T, en se basant sur l'attention de personnes qui voient et des descriptions. Des tests sont nécessaires pour établir l'efficacité des représentations pour un public inclusif. Il sera notamment intéressant de voir les méthodes à privilégier pour évaluer les effets sonores générées automatiquement. La transmission de l'information par stimulus audio peut passer par des informations verbales ou non verbales (i.e., sonification). Pour des œuvres d'art, la sonification peut passer par la conversion de propriétés colorimétriques en propriétés du son (d'instrument/de mélodie [11]), ou encore la spatialisation d'audiodescription [12]. Plus récemment, des modèles d'apprentissage profond sont étudiés pour recommander des sons pouvant correspondre à des images. Kim *et al.* [13] propose d'utiliser un modèle faiblement supervisé à partir de Resnet101 et entraîné à partir de vidéos de publicités pour recommander des musiques dont l'ambiance sonore correspondrait à des images de peintures. L'utilisation de modèle de diffusion latente pour la génération d'audio de type effets sonores, vers des audios serait une piste intéressante pour faciliter l'intégration cognitive d'informations multimodales. Récemment, Liu *et al.* [14] proposent le modèle AudioLDM se basant sur des prompts pour générer des sons. Ce modèle est pré-entraîné sur plusieurs jeux de données (Audioset, Audiocaps, librairie d'effets sonores de la BBC, Freesound). Les textes en entrée nécessitent d'être assez bien décrits pour générer des effets sonores correspondant aux descriptions données.

Références

[1] L. Smith, J. Smith et P. Tinio, «Time Spent Viewing art and Reading Labels,» *Psychology of Aesthetics, Creativity, and the Arts,* vol. 11, pp. 77-85, 2017, DOI: 10.1037/aca0000049.

[2] J. Smith et . L. Smith, «Spending Time on Art,» *Empirical Studies of the Arts,* vol. 19, p. 229–236, 2001, DOI: 10.2190/5MQM-59JH-X21R-JN5J.

[3] A. Yarbus, Eye movements and vision, New York: PLENUM PRESS, 1967.

[4] K. Reinholt, D. Guinness et S. K. Kane, «Eyedescribe: Combining eye gaze and speech to automatically create accessible touch screen artwork,» *Proceedings of the 2019 ACM International Conference on Interactive Surfaces and Spaces,* p. 101–112, 2019, DOI:10.1145/3343055.3359722.

[5] J. Pont-Tuset, J. Uijlings, S. Changpinyo, . R. Soricut et V. Ferrari , «Connecting vision and language with localized narratives,» *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings. Springer International Publishing.*, vol. 16, pp. 647-664, 2020, DOI: 10.1007/978-3-030-58558-7_38.

[6] S. Gay , E. Pissaloux, K. Romeo et N. T. Truong, «F2T: a novel force-feedback haptic architecture delivering 2D data to visually impaired people,» *IEEE Access,* vol. 9, pp. 94901-94911, 2021.

[7] T. P. Way et K. E. Barner, «Automatic visual to tactile translation. i. human factors, access methods and image manipulation,» *IEEE Transactions on rehabilitation engineering,* vol. 5, n° %11, p. 81–94, 1997.

[8] A. Abdusalomov, M. Mukhiddinov, O. Djuraev, U. Khamdamov et T. K. Whangbo, «Automatic salient object extraction based on locally adaptive thresholding to generate tactile graphics,» *MDPI Applied Sciences,* vol. 10, pp. 33-50, 2020.

[9] C. Rother, V. Kolmogorov et A. Blake, «"GrabCut": interactive foreground extraction using iterated graph cuts,» *ACM SIGGRAPH 2004 Papers,* p. 309–314, 2004, DOI: 978-1-4503-7823-9.

[10] S. Xie et Z. Tu, «Holistically-Nested Edge Detection,» *International Journal of Computer Vision,* vol. 125, pp. 3-18, DOI: 10.1007/s11263-017-1004-z, 2017.

[11] J.-D. Cho, J. Jeong , J.-H. Kim et H. Lee, «Sound Coding Color to Improve Artwork Appreciation by People with Visual Impairments,» *MDPI Electronics,* 2020, DOI: 10.3390/electronics9111981.

[12] Y. Lee, C.-H. Lee et D.-C. Jun, «3D sound coding color for the visually impaired. Electronics,» *MDPI Electronics,* vol. 10, p. 1037, 2021, DOI: 10.3390/electronics10091037.

[13] Y. Kim, H. Jeong, J.-D. Cho et J. Shin, «Construction of a soundscape-based media art exhibition to improve user appreciation experience by using deep neural networks,» *MDPI Electronics,* vol. 10, p. 1170, 2021, DOI: 10.3390/electronics10101170.

[14] H. Liu, Z. Chen, Y. Yuan, X. Mei, X. Liu, D. Mandic, W. Wang et M. D. Plumbley, «Audioldm: Text-to-audio generation with latent diffusion models,» *arXiv preprint arXiv:2301.12503,* 2023, DOI: 10.48550/arXiv.2301.12503.

[15] M. . R. Morris, J. Johnson, C. L. Bennett et E. Cutrell, «Rich representations of visual content for screen reader users,» *Proc. of the 2018 CHI (conf. on Human Factors in Computing Systems), ACM,* pp. 1-11, 2018.