Reconnaissance des locuteurs à partir de la dynamique faciale

Mots clés : indentification à distance, reconnaissance faciale, leurrage, vidéo hyper-truquée, contre-attaque.

Contexte.

Habituellement, la reconnaissance faciale est réalisée à partir d'une image fixe, éventuellement extraite d'une vidéo. La reconnaissance faciale à distance via un téléphone intelligent est de plus en plus envisagée dans de nombreuses applications [1,2]. Afin de prévenir, ou tout au moins rendre plus difficile, les attaques de présentation (i.e. un usurpateur présente une photographie d'une tierce personne devant le capteur [3]), l'ANSSI préconise de réaliser l'authentification des personnes à partir de vidéos plutôt que d'images fixes [4]. Travailler à partir de vidéos plutôt que d'images fixes offre la possibilité d'utiliser la dynamique faciale des locuteurs (clignements des yeux, mouvements de la tête, mimiques faciales, sourire, expressions) et non plus simplement l'apparence. Nous passerions ainsi d'une biométrie physique (apparence du visage) à une biométrie comportementale (dynamique du visage). Cette transition rendra plus difficile le leurrage des systèmes même s'il faut d'ores et déjà tenir compte de l'émergence des vidéos hypertruquées (« deepfakes » [5] en anglais). On peut en effet imaginer dans un proche avenir la possibilité de générer en temps réel des deepfakes incluant quelques actions et expressions faciales éventuellement demandées par le système d'authentification en cours de session.

Objectif.

Cette thèse permettra d'identifier à la fois les mimiques faciales les plus significatives pour authentifier les individus (en complément de l'apparence) mais devra également préciser le niveau de difficulté pour les reproduire avec réalisme de manière artificielle afin de prévenir les attaques de présentation en vidéo. Il est important de distinguer ici si une attaque a pour objectif de leurrer (visuellement) un opérateur humain, ou (numériquement) un algorithme de reconnaissance faciale basé sur l'intelligence artificielle, ou bien encore les deux à la fois. Ces travaux ont pour le moment comme objectif, non pas de remplacer, mais d'apporter une aide aux opérateurs humains qui doivent décider si oui ou non ils sont en présence d'une tentative usurpation d'identité.

Organisation des travaux.

La première année, le doctorant devra se familiariser avec à la fois les générateurs de vidéos hyperréalistes et les détecteurs.

Il faut admettre qu'au jour d'aujourd'hui, les outils qui existent pour **créer** automatiquement des deepfakes ont des performances limitées et ont encore du mal à duper un observateur humain [6]. Mais les progrès sont extrêmement rapides. Les vidéos hyper-réalistes comme celles de Tom Cruise (voir illustration 1 ci-après) ont nécessité un acteur pour créer les séquences de référence et des heures de post traitements manuels pour obtenir un rendu réaliste [7]. Parmi les logiciels libres et automatiques existants, on peut citer DeepFaceLab, FaceSwap, First Order Motion Model [8] (voir illustration 2 ci-après). Ces outils bien qu'imparfait, constituent néanmoins un bon point de départ pour générer des deepfakes.

Illustration 1

Une vidéo hyper-truquée très réaliste de Tom Cruise (à droite). Le visage de Tom Cruise a été échangé avec le visage de l'homme à droite. La vidéo complète est visible à :
https://www.youtube.com/watch?v=wq-kmFCrF5Q

Illustration 2

A partir d'une seule image cible (à gauche) et d'une vidéo source (au centre), il est possible de créer une fausse vidéo (à droite) similaire à la vidéo source mais cette fois avec le visage de la personne présente dans l'image cible.

En ce qui concerne la **détection**, nous prendrons comme point de départ, les résultats du challenge DFDC (Deepfake Detection Challenge) [9]. Il s'agit de l'une des actions les plus importante dans le domaine, organisé par Facebook et Microsoft et la participation de nombreuses universités. La meilleure solution atteint un taux de détection de 82%. Mais ces performances diminuent à 65% lorsque l'algorithme est testé sur d'autres bases incluant des deepfakes créés avec d'autres générateurs que ceux utilisés dans le cadre du challenge (et donc connus des algorithmes d'apprentissage). Les meilleures performances ont été obtenues à partir de méthodes basées essentiellement sur des structures *EfficientNet*. Egalement, l'augmentation de données pour l'apprentissage obtenue en masquant des zones d'intérêts a été largement utilisée.

La méthodologie retenue consistera à des allers-retours entre détection et création (principe de base des **GANs** [10]). Plus on détectera les deepfakes, plus nous devrons améliorer la qualité la qualité des vidéos générées, et ainsi de suite. Une originalité de la thèse consistera à comparer les erreurs de détection entre observateurs humaines et algorithmes automatiques. Ainsi, nous resterons parfaitement à jour par rapport aux dernières évolution des deepfakes, ce qui permettra en même temps de faire évoluer les outils de détection.

Comme aussi bien les générateurs que les détecteurs utilisent l'intelligence artificielle et les réseaux profonds, il se pose la question de l'**explicabilité** des résultats, comme par exemple en reconnaissance faciale [11], et comme dans beaucoup d'autres domaines de la vision par ordinateur basés sur l'intelligence artificielle. Nous analyserons par exemple les performances des détecteurs par sous catégories de population par rapport au genre ou à l'âge.

Même si la première année sera essentiellement consacrée à la prise en main des outils d'intelligence artificielle pour la création et la détection de « deepfakes ». On peut envisager des contributions techniques pour améliorer aussi bien la génération que la détection de « deepfakes ». Il est également attendu des résultats en ce qui concerne les liens entre qualité objective (duper un algorithme automatique) et qualité subjective (duper un observateur humain) des deepfakes. Enfin, ces travaux devront apporter également des éléments en termes d'explicabilité dans les performances de détection des deepfakes.

En **deuxième année**, les travaux auront pour principal objectif de définir les actions (**expressions faciales**) qui permettront une détection plus aisée des vidéos hyper truquées face aux vidéo authentiques. L'enjeu est extrêmement important car il s'agit ici de savoir si son interlocuteur est une vraie personne bien présente devant son téléphone intelligent ou son ordinateur et non pas une vidéo artificielle générée en temps réel [12]. L'idée sera de demander à l'utilisateur de réaliser une expression particulière bien identifiée en temps réel afin de sécuriser la reconnaissance faciale de l'utilisateur.

Une piste à explorer sera basée sur les **micro-expressions**. Les capteurs des téléphones intelligents capturent de plus en plus d'informations (en termes de résolutions et définitions) ; plus que ce que les humains peuvent percevoir. Les progrès des capteurs sont donc désormais surtout utiles pour les algorithmes d'analyse d'images. On peut citer ici la capacité des machines à estimer le rythme cardiaque par exemple en observant les variations de couleur dans le visage (variations imperceptibles par un humain) [13]. On peut également citer également les micro-expressions. Toutes ces informations « cachées » ne sont pas générées actuellement dans les « deepfakes » et donc devraient pouvoir être utilisées pour reconnaitre un utilisateur authentique d'un utilisateur simulé.

Ensuite, il faudra faire le lien entre la vidéo entrante (supposée authentique) et le ou les algorithmes de reconnaissance faciale. Dans le schéma actuel, nous avons deux blocs séparés, en séquence : le premier doit vérifier qu'il s'agit d'un **véritable locuteur** et le second doit vérifier **l'identité de la personne**. On peut envisager un **module conjoint** (par analogie avec les codages source et canal) qui intègrerait les deux objectifs de manière optimisée.

Le contenu exact de la **troisième année** sera défini en fonction des résultats obtenues les deux premières années, de l'avancement de l'état de l'art dans le domaine (assez difficile à prédire car extrêmement rapide), et de l'arrivée éventuelle de nouvelles réglementations.

On peut d'ores et déjà cependant prédire que le traitement des contenus multimédia par l'intelligence artificielle vont se multiplier (y compris directement au niveau capteurs). Ce n'est pas encore le cas aujourd'hui pour la compression comme MPEG encore basée sur des techniques classiques (transformée en cosinus et appariements de blocs) mais la situation évolue rapidement, e.g. JPEG AI [14]. Ainsi un nouveau défi va donc se poser dans un avenir assez proche. Il s'agira de savoir déterminer quel contenu a été traité par apprentissage profond à des fins malveillantes et quel contenu a été traité à des fins bienveillantes. Certains traitements bienveillants pouvant rendre plus difficile la détection de traitements malveillants car tous basés sur des technologies d'apprentissage profond, ou au contraire augmenter le nombre de faux positifs dans la détection de deepfakes.

La fin de thèse inclut classiquement un double objectif : intégrer des résultats de recherche dans des produits (ou à défaut des démonstrateurs) de DOCAPOST, rédiger le rapport de thèse et préparer la soutenance.

*La **stratégie de publications** sera définie par l'encadrant académique en liaison avec l'encadrant industriel afin d'inclure également la possibilité de déposer un ou plusieurs **brevets** au cours de la thèse. Nous prévoyons de commencer à publier en fin de première année ou début de seconde selon l'avancement des travaux. On vise des conférences majeures en traitements d'image comme IEEE ICIP (Int. Conference on Image Processing), des sessions spéciales sur le leurrage en reconnaissance faciale dans des conférences biométriques comme ICB (International Conference on Biometrics) ou bien encore des conférences sur les INFOX incluant des sessions sur deepfakes.*

Frédéric Dufaux

Bibliographie.

[1] J. Catalina Rodriguez (Gemalto), 01/04/19, https://www.journaldunet.com/solutions/reseau-social-d-entreprise/1422924-6-bonnes-raisons-de-recourir-a-la-reconnaissance-faciale/

[2] Thales, « Reconnaissance faciale : 7 tendances à suivre pour 2021 » https://www.thalesgroup.com/fr/europe/france/dis/gouvernement/biometrie/reconnaissance-faciale

[3] M De Marsico, M Nappi, D Riccio, JL Dugelay, « Moving face spoofing detection via 3D projective invariants", 2012 5th IAPR International Conference on Biometrics (ICB), 73-78.

[4] SGDSN/ANSS, Prestataire de vérification d'identité à distance – Référentiel d'exigences, I, 19, Novembre 2020.

[5] Verdoliva L. Media forensics and deepfakes: an overview. IEEE Journal of Selected Topics in Signal Processing. 2020 Jun 12;14(5):910-32.

[6] M. Tual (Le Monde), "On a essayé de fabriquer un deepfake (et on est passé à autre chose", 7 janvier 2020.

[7]J. Kahn (Tech News), Why deepfake creators love Tom Cruise, March 1, 2021.

[8] A. Siarohin et al., First Order Motion Model for Image Animation, NeurIPS (2019).

[9] B. Dolhansky et al. (Facebook AI), "The Deepfake Fetection Challenge (DFDC) Dataset, arXiv:2006.07397, 28 Oct. 2020.

[10] Generative adversarial network (GAN)
https://en.wikipedia.org/wiki/Generative_adversarial_network

[11] CHIST-ERA XAIface project, site <web: https://www.chistera.eu/projects/xaiface

[12] Maxime Recoquillé (L'EXPRESS) « Je peux vous appeler avec votre propre visage" : la folle avancée des deepfakes inquiète», 19/03/2021.

[13] P. V. Rouast et al., "Remote heart rate measurement using low-cost RGB face video: a technical literature review", DOI 10.1007/s11704-016-6243-6.

[14] J. Ascenso et al., Learning-based image coding: early solutions reviewing and subjective quality evaluation, SPIE Photonics Europe - Optics, Photonics and Digital Technologies for Imaging Applications VI, vol. 11353, Online, 2020.