Robustesse aux attaques en Authentification Digitale par Apprentissage Profond

Introduction.

L'authentification des personnes devient de plus en plus automatique et digitale dans nos sociétés. Les systèmes de passage de frontière tels que PARAFE ou les applications sur téléphone portable de vérification de l'identité d'une personne par la biométrie sont courantes. En particulier, la reconnaissance faciale est de plus en plus utilisée pour identifier les personnes afin d'accéder à un lieu ou une ressource. L'indentification peut se faire à distance via un smartphone ou un ordinateur. Il s'agit alors souvent de comparer un document d'identité présenté à la caméra avec un selfie du porteur du document. Les performances en termes de taux de reconnaissance ont fortement progressé ces dernières années grâce essentiellement à l'évolution des capteurs et les récentes avancées en apprentissage profond.

Malheureusement la robustesse de ces outils de reconnaissance faciale ou de vérification de documents n'a pas encore été suffisamment étudiée et des failles existent. Deux grandes catégories d'attaques ont été identifiées :

- Les attaques situées au niveau de l'acquisition (ou attaques physiques),
- Les attaques situées au niveau des images (ou attaques logicielles).

Les attaques liées à l'acquisition sont appelées « Presentation Attacks ». Il s'agit de présenter une photographie ou porter un masque d'une autre personne. Il peut s'agir également de présenter un document d'identité falsifié.

Les attaques logicielles au niveau des images portent sur les mêmes ressources mais utilisent une approche de type Man-in-the-Middle (MITM) pour détourner le flux de données d'un capteur tel que la caméra à travers un algorithme de transformation capable de fournir une vision faussée de la scène capturée. Le propriétaire d'un smartphone et qui souhaite tromper une application d'identification peut installer des kits fournissant la fonction MITM dans ce but. Cette pratique est déjà d'un usage courant pour leurrer les positions GPS. L'attaque sur authentification consiste alors à retoucher les images acquises afin de perturber le système avec l'un des objectifs suivants :

- soit le « Leurrage » : usurper l'identité de quelqu'un ou usurper ses droits,
- soit l'« Evasion » : éviter de se faire détecter et/ou reconnaitre.

Ce sujet de recherches vise à détecter ces attaques au moyen de méthodes de traitements des images et principalement par de l'apprentissage profond. Notre approche consistera à la fois à mettre en œuvre des mécanismes de détection sur des flux de données existants, mais aussi à créer des stratégies permettant de rendre ces détections plus faciles.

Des images fixes à la vidéo.

Jusqu'à présent, les travaux ont surtout porté sur les images fixes aussi bien au niveau de la reconnaissance faciale que des attaques possibles et des contre-attaques. Comme déjà indiqué les technologies de reconnaissance faciales à partir d'une image fixe marchent très bien dans les environnements contrôlés. Lorsque les conditions sont plus difficiles voire adverses, les performances chutent sensiblement et les systèmes sont exposés aux attaques : les attaques de présentations qui consistent à montrer une photographie ou à porter un masque d'une tierce personne (des équipes ont pas exemple réussi à piéger le déverrouillage de l'iPhone X) ; les attaques d'évasion qui consiste plus particulièrement à ce qu'un système ne puisse reconnaitre un individu dans un lieu public alors qu'il figure dans une liste noire. Une des dernières attaques connues est intitulée « face morphing ». Elle permet la construction de visages artificiels

situés entre deux visages réels, autorisant la reconnaissance biométrique de deux personnes (suffisamment proches) à partir de la même image [1].

Une parade à ces vulnérabilités est d'utiliser la vidéo plutôt que les images fixes [2]. Aujourd'hui certains opérateurs d'identification utilisent des interviews via des logiciels de conférence vidéo pour authentifier une personne. Malheureusement de nouvelles attaques existent d'ores et déjà également en vidéo et on peut s'attendre à ce que ces attaques se perfectionnent dans les années à venir. Les flux vidéo transmis peuvent aujourd'hui être manipulés en temps réel de sorte à faire apparaitre la réactivité faciale d'un faussaire sur le visage d'une autre personne [3]. Certains logiciels librement diffusés permettent L'échange de visage (FaceSwapping) [4]. D'autres permettent le même type d'illusion au niveau des mouvements corporels [5]. Il est donc primordial de commencer dès à présent à travailler aux ripostes. Il sera également important d'anticiper l'évolution rapides des capteurs vidéos.

Objectifs de la thèse

L'enjeu de cette thèse est donc de développer des méthodologies permettant de garantir l'authenticité d'une vidéo d'identification et en cas de non-authenticité de détecter la nature des falsifications.

Ces méthodologies peuvent être passives (analyse des propriétés des images et des vidéos) ou actives (induction de clignement des yeux, de rotation de la tête, de mouvement du téléphone, etc.). On s'intéressera donc aux deux ainsi qu'à de possibles combinaisons.

Une partie de la thèse consistera donc à définir des ripostes et des protocoles sur la captation des flux, sur l'interaction avec les personnes, sur l'usage de la caméra, sur l'emploi de tests, etc. qui permettront d'augmenter la confiance associée à un

flux vidéo. Il sera primordial de considérer des éléments technologiques de mis en œuvre pratiques comme le temps de calcul ou l'acquisition à partir d'un téléphone intelligent ; et des éléments légaux liés au respect de la vie privée.

État de l'Art

De nombreux travaux existent concernant les méthodes passives, principalement autour de la communauté de l'Image Forensics[7,8,9] qui recherche les anormalités au sein des images ou des flux pour en déduire la localisation des manipulations. Mais ces travaux concernent les images et les vidéos en général et sont faiblement représentatives de la problématique de l'usurpation d'identité, alors que cette thèse se focalisera sur les images et les vidéos de visage dédiées à l'identification des personnes. Les modifications sont détectées grâce à des incohérences ou des anormalités estimées sur les points de l'image, des incohérences sur le bruit des capteurs, des re-compressions[10,15], des recopies internes[11], externes[12], des incohérences en termes d'illumination, ou des contours[13, 14]. Plusieurs challenges technologiques ont été lancés par la DARPA, l'IEEE et le NIST aux Etats-Unis ou encore la DGA en France (dont un DEFALS auquel EURECOM & SURYS participent) pour mesurer l'efficacité de ce type de méthode. Il est à noter que des progrès significatifs ont été récemment réalisés grâce aux techniques d'apprentissage profond. La détection passive peut également s'appuyer sur les particularités des attaques, comme par exemple ce qui est réalisé lors d'un morphing entre deux images[16], ou sur l'historique des opérations portant sur celles-ci[17].

Concernant les méthodes actives, là encore de nombreuses stratégies ont été inventées pour prévenir les « attaques de présentation » en biométrie : depuis les demandes de clignements des yeux, la déformation de la bouche lors de production de parole, la rotation de la tête, l'utilisation des mains pour masquer telle ou telle partie du visage jusqu'au tests interactifs dirigé par un humain qui cherche à déstabiliser le sujet pour l'amener à des situations non prévues par un usurpateur.

Cependant l'efficacité de ces ripostes commencent à être remis en cause par les progrès des technologies d'inpainting par deep learning qui ont amené une très forte crédibilité à des images de synthèse produites en temps réel sur la base de quelques photos de la personne dont on usurpe l'identité et d'un flux vidéo de l'usurpateur répondant (potentiellement) à toutes les sollicitations des tests précédents [5]. Le Face spoofing peut cependant être détecter dans les flux vidéos en s'attachant aux caractéristiques connues des images traitées, comme par exemple les particularités 3D d'un visage[18].

Directions d'exploration

Une première direction consiste à exploiter les particularités des images biométriques en tant que sous-ensemble des photographies pour rendre plus efficace les méthodes de détection de manipulations (image forensics, imagerie légale). Le fait de connaitre par exemple la présence de traits spécifiques sur un visage (décrochement et angle du nez, position des commissures des lèvres, etc…) permet de concevoir des tests spécifiques qui seront impactés par les méthodes visant à altérer la biométrie. Le nombre et la variété de ces caractéristiques nous amène à l'emploi de méthodes d'apprentissage profond pour éviter d'avoir à concevoir des algorithmes spécifiques.

Une deuxième direction consiste à intégrer la problématique de la falsification dans le cadre non d'une image statique, mais dans celui d'un flux vidéo et même interactif. Cette dimension rend immédiatement largement plus difficile le travail de l'imposteur et nous ouvre de nombreux champs d'authentification, comme la perception en trois dimensions du visage (ce qui peut être conforté par des technologies comme les capteurs IR/thermiques actuellement en forte progression chez les équipementiers).

La dimension interactivité de l'authentification amène également à la problématique de la conception de nouveaux jeux de tests, au-delà des classiques clignement des yeux et rotation de tête, qui puissent permettre une exploitation idéale des deux premières directions. Cette interactivité doit être prise au sens large, comme par exemple par la variation codée de la luminosité de l'écran d'un

smartphone lors d’un selfie pour assurer l’intégrité de la séquence capturée. La conception de ces tests constitue la troisième direction de cette thèse.

Points novateurs de la thèse

- Explorer un domaine nouveau à l’intersection de deux domaines aujourd’hui indépendants : imagerie légale (i.e. intégrité des images) et biométrie (reconnaissance faciale) ;
- Explorer la problématique en vidéo plutôt qu’en image fixe ;
- Explorer la problématique sous l’angle de l’interactivité ;
- Traiter la problèmatique dans une approche d’Apprentissage Profond ;
- Anticiper l’apparition de nouvelles modalités de contrôle (i.e. capteur thermique) ;

Bibliographie.

[1] Ferrara, M., Franco, A., & Maltoni, D. (2014, September). The magic passport. In *Biometrics (IJCB), 2014 IEEE International Joint Conference on* (pp. 1-7). IEEE.

[2] F Matta, JL Dugelay. Person recognition using facial video information: A state of the art. In *Journal of Visual Languages & Computing* 20 (3), 180-187

[3] Thies, J., Zollhofer, M., Stamminger, M., Theobalt, C., & Nießner, M. (2016). Face2face: Real-time face capture and reenactment of rgb videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 2387-2395).

[4] Bitouk, D., Kumar, N., Dhillon, S., Belhumeur, P., & Nayar, S. K. (2008, August). Face swapping: automatically replacing faces in photographs. In *ACM Transactions on Graphics* (TOG) (Vol. 27, No. 3, p. 39). ACM.

[5] Chan, C., Ginosar, S., Zhou, T., & Efros, A. A. (2018). Everybody dance now.

[6] Güera, D., & Delp, E. J. Deepfake Video Detection Using Recurrent Neural Networks.

[7]JA Redi, W Taktak, JL Dugelay, Digital image forensics: a booklet for beginners, Multimedia Tools and Applications 51 (1), 133-162

[8] Fridrich, J., Soukal, D., et Lukas, J., "Detection of Copy-Move Forgery in Digital Images," Digital Forensic Research Workshop, August 2003.

[9] Popescu, A.C., "Statistical tools for digital image forensics," Ph.D. Thesis, Department of Computer Science, Dartmouth College, Hanover, USA, 2005

[10] Li, W., Yuan, Y., et Yu, N., "Passive detection of doctored JPEG image via block artifact grid extraction," Signal Processing, vol. 89, no. 9, pp. 1821-1829, September 2009.

[11] Pan, X., Lyu, S., "Region duplication detection using image feature matching" IEEE Transactions on Information Forensics and Security, vol. 5, n° 4, p. 857-867, December 2012.

[12] Kakar, P., N Sudha, N., "Exposing Postprocessed Copy-Paste Forgeries Through Transform-Invariant Features" IEEE Transactions on Information Forensics and Security, vol. 7, n° 3, p. 1018-1028, June 2012.

[13] Johnson, M., et Farid, H., "Exposing digital forgeries through chromatic aberration," ACM Workshop on Multimedia and Security (MMSEC), pp. 48-55, September 2006.

[14] Gloe, T., Winkler, A., et Borowka, K., "Efficient estimation and large-scale evaluation of lateral chromatic aberration for digital image forensics," IS&T/SPIE conf. on Electronic Imaging, Security and Forensics of Multimedia Contents XII, pp. 107-114, January 2010.

[15] Farid, H., "Exposing digital forgeries from JPEG ghosts", IEEE Transactions on Information Forensics and Security, vol. 4, no. 1, pp. 154-160, March 2009.

[16] R Raghavendra, KB Raja, C Busch, Detecting morphed face images, Biometrics Theory, Applications and Systems (BTAS), 2016 IEEE 8th

[17] N Kose, JL Dugelay, Classification of captured and recaptured images to detect photograph spoofing, Informatics, Electronics & Vision (ICIEV), 2012 International Conference on Informatics, Electronics &Vision

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