Vers un alignement spatio-temporel du visage en conditions non contrôlées

Romain Belmonte, Nacim Ihaddadene, Pierre Tirilly, Ioan Marius Bilasco, Chaabane Djeraba

Vers un alignement spatio-temporel du visage en conditions non contrôlées

R. Belmonte[1,3,*] N. Ihaddadene[1,*] P. Tirilly[2,†] M. Bilasco[3,†] C. Djeraba[2,†]

[1] ISEN Lille, Yncréa Hauts-de-France, France
[*]{romain.belmonte, nacim.ihaddadene}@yncrea.fr

[2] Univ. Lille, CNRS, Centrale Lille, IMT Lille Douai, UMR 9189 - CRIStAL - Centre de Recherche en Informatique Signal et Automatique de Lille, F-59000 Lille, France

[3] Univ. Lille, CNRS, Centrale Lille, UMR 9189 - CRIStAL - Centre de Recherche en Informatique Signal et Automatique de Lille, F-59000 Lille, France
[†]{pierre.tirilly, marius.bilasco, chaabane.djeraba}@univ-lille1.fr

Résumé

L'alignement du visage est une tâche essentielle pour un grand nombre d'applications. Elle a pour objectif de localiser des points caractéristiques sur le visage, dans le but d'identifier sa structure géométrique. En conditions non contrôlées, les différentes variations pouvant intervenir dans le contexte visuel, associées à l'instabilité de la détection du visage, en font un problème difficile à résoudre. Bien que de nombreuses méthodes aient été proposées, leurs performances en présence de ces contraintes ne sont toujours pas satisfaisantes. Dans cet article, nous invitons à étudier l'alignement du visage à l'aide de séquences d'images et non d'images fixes, comme cela a pu être réalisé jusqu'à présent, et montrons l'importance de la prise en compte de l'information temporelle en conditions non contrôlées.

Mots clefs

Alignement du visage, approche temporelle, points caractéristiques, séquences d'images, conditions non contrôlées.

1 Introduction

Le problème de l'alignement du visage, aussi appelé localisation des points caractéristiques du visage, a suscité beaucoup d'intérêt et connu de rapide progrès ces dernières années [**?**]. Etant donné la position et la taille d'un visage, l'alignement, illustré en figure 1, consiste à déterminer la géométrie des composantes du visage contenant le plus d'information sémantique (pour ex., les yeux, le nez, la bouche). Cette capacité à modéliser les structures non rigides du visage est aujourd'hui exploitée dans divers domaines tels que l'analyse faciale (pour ex. identification, expressions), l'interaction homme-machine, ou le multimédia (pour ex. recherche, indexation). Toutefois, malgré le nombre important de méthodes présentes dans la littérature, les performances de l'alignement du visage en conditions non contrôlées restent limitées [**?**].

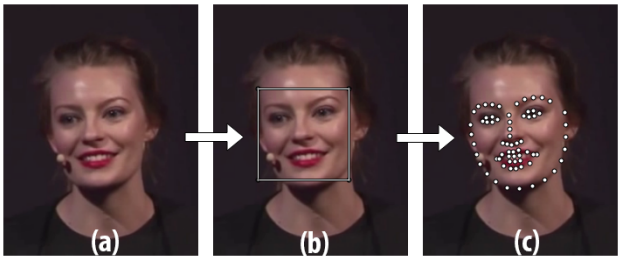

Figure 1 – *Processus d'alignement du visage : (a) image originale, (b) détection du visage, (c) alignement du visage. Images issues de 300VW [**?**].*

Encore aujourd'hui, ce problème continue d'être étudié à l'aide d'images fixes [**?**]. Pourtant, du fait de l'omniprésence des capteurs vidéos, la grande majorité des applications reposent sur des séquences d'images. De plus, bon nombre de tâches liées à l'analyse faciale ou plus largement, à l'analyse du comportement humain, ont su tirer profit de l'information temporelle [**?**, **?**]. Les premières synthèses sur l'alignement du visage ont déjà suggéré d'étudier le problème à l'aide de séquences d'images, mais sans toutefois avancer de réels arguments [**?**]. Notre motivation à travers cet article est de montrer que la prise en compte de l'information temporelle pour ce problème pourrait grandement contribuer à l'amélioration des performances en conditions non contrôlées.

Cet article est structuré de la manière suivante : dans la section 2, nous décrivons les raisons qui nous ont conduit à nous orienter vers une approche spatio-temporelle pour l'alignement du visage. Nous passons notamment en revue les solutions existantes et mettons en perspective les performances des méthodes les plus récentes. Nous consacrons ensuite la section 3 aux bénéfices de l'information temporelle en conditions non contrôlées. Enfin, nous concluons avec la section 4.

2 État des lieux

2.1 Jeux de données et métriques d'évaluation

Ces dernières années, de nombreux corpus pour l'alignement du visage ont été mis à la disposition de la communauté scientifique (c.f. Tableau 1). Les images incluses dans ces corpus ont été collectées sur des réseaux sociaux tels que Google, Flickr, ou Facebook, apportant ainsi plus de réalisme à l'étude du problème. L'annotation a été réalisée soit manuellement, soit de manière semi-automatique, avec parfois l'aide de la plate-forme Amazon Mechanical Turk. La qualité des annotations peut toutefois être assez variable [**?**, **?**, **?**]. Le schéma d'annotation utilisé (c.-à-d., position et nombre de points) peut également différer d'un jeu de données à un autre. À l'heure actuelle, c'est le schéma composé de 68 points [**?**], illustré en Figure 1, qui est le plus largement utilisé. Ce schéma ne permettant pas de conserver une correspondance pour la totalité des points lors de poses extrêmes, un schéma basé sur 39 points a récemment été proposé pour les visages de profil [**?**].

Type	Base de données	#Images	#Points
Statique	AFLW [**?**]	25.993	21
	AFW [**?**]	250	6
	HELEN [**?**]	2330	194
	iBug [**?**]	135	68
	LFPW [**?**]	1432	35
	COFW [**?**]	1007	29
	300W [**?**]	3837	68
	300W-LP [**?**]	61.225	68
	MENPO [**?**]	∼9000	68/39
Dynamique	300VW [**?**]	218.595	68

Tableau 1 – *Jeux de données capturés en conditions non contrôlées.*

L'évaluation d'une prédiction se fait généralement à l'aide de l'erreur quadratique moyenne normalisée par la distance interoculaire. Au delà de 8%, la prédiction est généralement considérée comme un échec. La normalisation par la distance interoculaire, bien que peu robuste aux poses extrêmes, est la plus répandue. D'autres normalisations sont parfois utilisées, comme notamment la diagonale de la fenêtre de détection. Sur un ensemble d'images, l'erreur moyenne est la métrique d'évaluation la plus simple et intuitive à calculer. Toutefois, elle peut être fortement impactée par la présence de quelques erreurs aberrantes. Une représentation graphique de la fonction de répartition de l'erreur est, de ce fait, de plus en plus souvent employée. Elle correspond à la proportion d'images pour laquelle l'erreur est inférieure ou égale à un certain seuil (par ex., 8%). L'aire sous la courbe et le taux d'échec, c'est-à-dire le pourcentage d'images pour lesquelles l'erreur est supérieure au seuil défini, sont parfois calculés à partir de cette représentation.

Aujourd'hui, du fait notamment de l'émergence des techniques d'apprentissage profond, il peut être nécessaire de disposer d'un grand nombre d'images annotées, ce que ne permettent pas forcément les jeux de données existants. Dans la littérature, différents procédés d'augmentation sont utilisés pour contourner ce problème. Certaines opérations peuvent être appliquées sur les images afin d'en générer de nouvelles (par ex., rotation, inversion horizontale, perturbation de la fenêtre de détection). D'autres procédés plus complexes tels que la génération d'images de synthèses commencent également à être employés [**?**].

Figure 2 – *Illustration des défis rencontrés en conditions non contrôlées : occultations (b), (d), (f), variations de pose (a), (e), éclairage (a), (b), expressions (c). Images issues de 300VW [**?**].*

Il est crucial d'avoir des données représentatives du problème afin de pouvoir y répondre. Les jeux de données statiques ne couvrent pas toutes les difficultés rencontrées par les applications, largement basées sur des séquences d'images. Les contraintes notamment liées au mouvement des personnes ou de la caméra ne sont actuellement pas considérées. Un corpus composé de séquences d'images capturées en conditions non contrôlées a toutefois récemment été publié par Shen et al. [**?**] (cf. Figure 2). Ces données, en plus d'être plus représentatives du problème, fournissent des éléments de réflexion en faveur de l'utilisation de l'information temporelle pour l'alignement du visage (cf. Section 3).

2.2 Synthèse des solutions

Dans la littérature se distinguent deux grandes catégories de méthodes pour localiser les points caractéristiques du visage. Il y a tout d'abord les méthodes dites génératives, qui s'appuient sur des modèles paramétriques conjoints d'apparence et de forme [**?**]. L'alignement est alors formulé comme un problème d'optimisation avec comme objectif de trouver les paramètres permettant de générer la meilleure instance possible du modèle pour un visage donné. L'apparence peut être représentée aussi bien de ma-

nière holistique que locale, à l'aide de régions d'intérêt centrées sur les points caractéristiques.

Il y a ensuite les méthodes dites discriminatives qui infèrent la position des points caractéristiques directement à partir de l'apparence du visage. Cela est rendu possible soit par l'apprentissage de détecteurs locaux indépendants ou de régresseurs pour chaque point caractéristique associés à un modèle de forme permettant de régulariser les prédictions [**?**], soit par l'apprentissage d'une ou plusieurs fonctions de régression vectorielles capablent d'inférer l'ensemble des points caractéristiques et de conserver implicitement une contrainte de forme [**?**, **?**, **?**, **?**]. Dans cette catégorie, les méthodes basées sur des techniques d'apprentissage profond (par ex., réseaux de neurones convolutionnels, auto-encodeurs) ont récemment permis d'améliorer de manière conséquente les performances en conditions non contrôlées grâce notamment à leur capacité à modéliser la non-linéarité et à apprendre des caractéristiques spécifiques au problème [**?**, **?**, **?**, **?**].

Bien que la plupart des méthodes s'attaquent de manière globale au problème, certaines se concentrent spécifiquement sur une difficulté [**?**, **?**, **?**]. Burgos-Artizzu et al. [**?**] se sont intéressés à modéliser explicitement les occultations et ont montré que cette information supplémentaire aidait à améliorer l'estimation de la position des points caractéristiques en conditions non contrôlées. L'apprentissage nécessite cependant un travail conséquent d'annotation des occultations. Zhu et al. [**?**] se sont eux focalisés sur les poses extrêmes et ont proposé d'inférer un modèle dense 3D plutôt qu'un modèle éparse 2D. Leur méthode est capable de gérer des variations de pose horizontale allant de -90˚à 90˚.

D'autres travaux suggèrent que l'alignement du visage ne doit pas être traité comme un problème indépendant et proposent d'apprendre conjointement différentes tâches connexes afin d'obtenir des gains de performances individuels [**?**, **?**]. Dans les travaux de Zhang et al. [**?**], l'alignement n'est pas appris de manière isolé mais conjointement avec différentes tâches connexes telles que l'estimation de la pose, du genre, des expressions faciales et de l'apparence des attributs faciaux. Ce type d'approche peut toutefois rendre l'étape d'apprentissage beaucoup plus complexe en raison des taux de convergence pouvant être variables d'une tâche à une autre.

Les performances des méthodes d'alignement du visage les plus récentes sont référencées dans le Tableau 2. Ces méthodes ont été évaluées sur le jeu de données 300W, composé de catégories de difficultés variables. La catégorie 300-A correspond à des images ne présentant pas de fortes contraintes. La catégorie 300-B contient des images plus complexes présentant de fortes variations de pose et d'expression, ainsi que des occultations. On constate que pour la catégorie B l'erreur moyenne vaut plus du double de celle obtenue sur la catégorie A. Les résultats toutes catégories confondues ne sont pas forcément pertinents du fait du déséquilibre entre les deux catégories (554 images pour la catégorie A contre 135 pour la catégorie B).

Méthode	**300W-A**	**300W-B**	**300W**
FPLL [**?**]	8.22	18.33	10.20
RCPR [**?**]	6.18	17.26	8.35
DRMF [**?**]	6.65	19.79	9.22
SDM [**?**]	5.57	15.40	7.50
ESR [**?**]	5.28	17.00	7.58
CFAN [**?**]	5.50	16.78	7.69
LBF [**?**]	4.95	11.98	6.32
RCR [**?**]	4.83	12.02	6.24
CFSS [**?**]	4.73	9.98	5.76
CMC-CNN [**?**]	4.91	12.03	6.30
RPPE [**?**]	5.50	11.57	6.69
3DDFA [**?**]	6.15	10.59	7.01
TCDCN [**?**]	4.80	8.60	5.54
RAR [**?**]	4.12	**8.35**	**4.94**
RCFA [**?**]	**4.03**	9.85	5.32
R-DSSD [**?**]	4.16	9.20	5.59

Tableau 2 – *Performances des méthodes récentes de la littérature. L'erreur quadratique normalisée moyenne est reportée.*

Malgré la quantité de méthodes proposées dans la littérature et les avancées majeures récentes, nous pouvons voir à travers ces résultats que les problèmes rencontrées en conditions non contrôlées sont encore loin d'être résolus. Du fait de leur influence significative sur l'apparence du visage, les variations de pose et les occultations font partie des défis les plus difficiles à relever. Nous montrons dans la Section 3 comment l'approche temporelle pourrait contribuer à résoudre ces problèmes.

3 Les bénéfices de l'information temporelle

3.1 Suivi non rigide

La détection du visage en conditions non contrôlées est un problème complexe à résoudre [**?**]. Étant donné son rôle pour l'alignement du visage, Yang et al. [**?**] se sont intéressés à la dépendance susceptible de s'exercer entre ces deux tâches. Leur étude a pu mettre en évidence une forte sensibilité de l'alignement à la détection. Ainsi, au-delà d'une incapacité à détecter un visage, d'autres facteurs tels que les variations d'échelle et de position de la fenêtre de détection peuvent venir perturber l'alignement.

Une solution pour s'abstraire de la dépendance à la détection est d'effectuer un suivi non rigide du visage. Shen et al. [**?**] ont récemment proposé une analyse comparative des méthodes actuelles de suivi non rigide spécifique au visage. La stratégie la plus populaire reste le suivi par détection, c'est-à-dire la détection du visage et son alignement sur chaque image de manière indépendante, sans tirer profit

des images adjacentes. Yang et al. [**?**] ont toutefois montré des résultats supérieurs à cette stratégie en proposant une régression en cascade spatio-temporelle. Leur méthode consiste à initialiser la forme sur l'image courante à partir des paramètres de similitude de l'image précédente. Elle intègre également un mécanisme de ré-initialisation basé sur la qualité de la prédiction afin d'éviter toute dérive de l'alignement. Comparée au suivi par détection, la régression en cascade spatio-temporelle permet de réduire considérablement le taux d'échec tout en améliorant les performances (cf. Tableau **??**).

Une alternative au suivi par détection consiste à utiliser, en substitution à la détection, un algorithme de suivi générique (c.-à-d. rigide). L'un des avantages des algorithmes de suivi génériques est qu'ils sont capables de tenir compte de certaines variations d'apparence de l'objet cible durant le suivi [**?**]. Chrysos et al. [**?**] ont évalué cette stratégie et l'ont comparé au suivi par détection. De manière générale, le suivi générique permet d'être plus robuste aux contraintes rencontrées en conditions non contrôlées. Il est cependant probable que, tout comme avec la détection, l'alignement soit sensible aux variations de la fenêtre de suivi.

Tenir compte des variations d'apparence sans passer par un algorithme de suivi générique devient alors judicieux. Sanchez et al. [**?**] ont proposé une régression en cascade continue incrémentale. Contrairement à Yang et al. [**?**] qui conservent un modèle générique une fois l'apprentissage effectué, ici un modèle pré-entraîné est mis à jour en ligne afin de devenir spécifique à la personne au cours du suivi. Ce genre d'approche montre de meilleurs résultats qu'avec un modèle générique mais n'exploite cependant que les variations d'apparence. D'autres informations telles que la trajectoire des points caractéristiques à travers la séquence d'images semblent pertinentes à considérer [**?**].

3.2 Contraintes supplémentaires

Qu'elle soit explicite ou implicite, la contrainte de forme présente dans la plupart des méthodes d'alignement est un élément crucial pour obtenir de bonnes performances en conditions non contrôlées. Dans des séquences d'images, une contrainte supplémentaire peut être appliquée sur la trajectoire des points caractéristiques. Peng et al. [**?**] ont utilisé un auto-encodeur récurrent afin d'exploiter les caractéristiques dynamiques du visage. Ils ont comparé une version récurrente et une non récurrente de leur auto-encodeur et ont montré que l'apprentissage récurrent améliorait, d'une part, la stabilité des prédictions et, d'autre part, la robustesse aux occultations, variations de pose et d'expression. Ils supposent notamment que le réseau de neurones récurrent a permis l'apprentissage de motifs de trajectoires, mais ne l'ont cependant pas démontré.

Une autre option pour tenir compte de la dynamique faciale est l'utilisation de filtres bayésien tels que les filtres de Kalman ou les filtres particulaires. Gu et al. [**?**] ont cependant mis en évidence le gain marginal apporté par ces approches et on montré la supérioté des réseaux de neurones récurrents pour l'analyse dynamique du visage. Les performances de leur méthode sont référencées dans le Tableau **??**. La prise en compte de contraintes supplémentaires montre une réduction notable de l'erreur moyenne de plus de 1% comparée à une méthode statique actuelle.

Les réseaux de neurones récurrents, bien qu'avantageux, ne sont capables de caractériser que le mouvement global. Dans d'autres tâches connexes, le mouvement local (c.-à-d. sur quelques trames) parfois associé au mouvement global a toutefois su montrer son intérêt [**?**, **?**] et pourrait être tout aussi bénéfique à l'alignement du visage.

4 Conclusion

Dans ce papier nous avons présenté une synthèse des travaux sur l'alignement du visage en conditions non contrôlées. Malgré des applications majoritairement basées sur des séquences d'images, ce problème est depuis plusieurs décennies étudié à partir d'images fixes. Un nombre impressionnant de méthodes sont proposées dans la littérature. Toutefois leurs performances en conditions non contrôlées ne sont toujours pas satisfaisantes du fait notamment des nombreuses variations pouvant intervenir dans le contexte visuel. Nous avons montré qu'étudier le problème à l'aide de séquences d'images pouvait grandement contribuer à pallier ces difficultés en plus d'améliorer la cohérence avec les applications. Nous pensons que l'ajout de contraintes temporelles est une voie intéressante à suivre.

Références

[1] Xin Jin et Xiaoyang Tan. Face alignment in-the-wild : A survey. *arXiv preprint arXiv :1608.04188*, 2016.

[2] Christos Sagonas, Epameinondas Antonakos, Georgios Tzimiropoulos, Stefanos Zafeiriou, et Maja Pantic. 300 faces in-the-wild challenge : Database and results. *Image and Vision Computing*, 47 :3–18, 2016.

[3] Jie Shen, Stefanos Zafeiriou, Grigoris G Chrysos, Jean Kossaifi, Georgios Tzimiropoulos, et Maja Pantic. The first facial landmark tracking in-the-wild challenge : Benchmark and results. Dans *ICCV Workshops*, pages 1003–1011. IEEE, 2015.

[4] Karen Simonyan et Andrew Zisserman. Two-stream convolutional networks for action recognition in videos. Dans *Advances in Neural Information Processing Systems*, pages 568–576, 2014.

[5] Yin Fan, Xiangju Lu, Dian Li, et Yuanliu Liu. Video-based emotion recognition using cnn-rnn and c3d hybrid networks. Dans *ICMI*, pages 445–450. ACM, 2016.

[6] Oya Çeliktutan, Sezer Ulukaya, et Bülent Sankur. A comparative study of face landmarking techniques. *EURASIP Journal on Image and Video Processing*, 2013(1) :13, 2013.

Méthode	Catégorie 1		Catégorie 2		Catégorie 3	
	AUC	FR(%)	AUC	FR(%)	AUC	FR(%)
HyperFace [**?**]	0.642	5.56	0.662	0.68	0.563	7.23
Yang et al. [**?**]	**0.791**	**2.400**	**0.788**	**0.322**	**0.710**	**4.461**
Uricar et Franc [**?**]	0.657	7.622	0.677	4.131	0.574	7.957
Xiao et al. [**?**]	0.760	5.899	0.782	3.845	0.695	7.379
Rajamanoharan et Cootes [**?**]	0.735	6.557	0.717	3.906	0.659	8.289
Wu et Ji [**?**]	0.674	13.925	0.732	5.601	0.602	13.161

Tableau 3 – *Comparaison des méthodes de l'analyse comparative de Shen et al. [**?**] avec Hyperface [**?**], méthode multitâche statique, sur les 3 catégories de 300VW. L'aire sous la courbe (AUC) et le taux d'échec (FR) sont reportés.*

Approche	**Méthode**	**300VW**
Statique	TCDCN [**?**]	7.59
Dynamique	RED-NET [**?**]	6.25
	Gu et al. [**?**]	**6.16**

Tableau 4 – *Comparaison de TCDCN [**?**], méthode multitâche statique, avec 2 méthodes dynamiques basées sur des réseaux de neurones récurrents [**?**, **?**]. L'erreur quadratique normalisée moyenne est reportée.*

[7] Adrian Bulat et Georgios Tzimiropoulos. How far are we from solving the 2d & 3d face alignment problem ?(and a dataset of 230,000 3d facial landmarks). *arXiv preprint arXiv :1703.07332*, 2017.

[8] Ralph Gross, Iain Matthews, Jeffrey Cohn, Takeo Kanade, et Simon Baker. Multi-pie. *Image and Vision Computing*, 28(5) :807–813, 2010.

[9] S Zaferiou. The menpo facial landmark localisation challenge. Dans *CVPR Workshops*, volume 1, 2017.

[10] Martin Köstinger, Paul Wohlhart, Peter M Roth, et Horst Bischof. Annotated facial landmarks in the wild : A large-scale, real-world database for facial landmark localization. Dans *ICCV Workshops)*, pages 2144–2151. IEEE, 2011.

[11] Xiangxin Zhu et Deva Ramanan. Face detection, pose estimation, and landmark localization in the wild. Dans *CVPR*, pages 2879–2886. IEEE, 2012.

[12] Vuong Le, Jonathan Brandt, Zhe Lin, Lubomir Bourdev, et Thomas Huang. Interactive facial feature localization. *ECCV*, pages 679–692, 2012.

[13] Christos Sagonas, Georgios Tzimiropoulos, Stefanos Zafeiriou, et Maja Pantic. A semi-automatic methodology for facial landmark annotation. Dans *CVPR Workshops*, pages 896–903, 2013.

[14] Peter N Belhumeur, David W Jacobs, David J Kriegman, et Neeraj Kumar. Localizing parts of faces using a consensus of exemplars. *IEEE Transactions on Pattern Analysis and Machine intelligence*, 35(12) :2930–2940, 2013.

[15] Xavier P Burgos-Artizzu, Pietro Perona, et Piotr Dollár. Robust face landmark estimation under occlusion. Dans *ICCV*, pages 1513–1520, 2013.

[16] Xiangyu Zhu, Zhen Lei, Xiaoming Liu, Hailin Shi, et Stan Z Li. Face alignment across large poses : A 3d solution. Dans *CVPR*, pages 146–155, 2016.

[17] Timothy F. Cootes, Gareth J. Edwards, et Christopher J. Taylor. Active appearance models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23(6) :681–685, 2001.

[18] Jason M Saragih, Simon Lucey, et Jeffrey F Cohn. Deformable model fitting by regularized landmark mean-shift. *International Journal of Computer Vision*, 91(2) :200–215, 2011.

[19] Xuehan Xiong et Fernando De la Torre. Supervised descent method and its applications to face alignment. Dans *CVPR*, pages 532–539, 2013.

[20] Xudong Cao, Yichen Wei, Fang Wen, et Jian Sun. Face alignment by explicit shape regression. *International Journal of Computer Vision*, 107(2) :177–190, 2014.

[21] Shaoqing Ren, Xudong Cao, Yichen Wei, et Jian Sun. Face alignment at 3000 fps via regressing local binary features. Dans *CVPR*, pages 1685–1692, 2014.

[22] Shizhan Zhu, Cheng Li, Chen Change Loy, et Xiaoou Tang. Face alignment by coarse-to-fine shape searching. Dans *CVPR*, pages 4998–5006, 2015.

[23] Qiqi Hou, Jinjun Wang, Lele Cheng, et Yihong Gong. Facial landmark detection via cascade multi-channel convolutional neural network. Dans *ICIP*, pages 1800–1804. IEEE, 2015.

[24] Shengtao Xiao, Jiashi Feng, Junliang Xing, Hanjiang Lai, Shuicheng Yan, et Ashraf Kassim. Robust facial landmark detection via recurrent attentive-refinement networks. Dans *ECCV*, pages 57–72. Springer, 2016.

[25] Wei Wang, Sergey Tulyakov, et Nicu Sebe. Recurrent convolutional face alignment. Dans *ACCV*, pages 104–120. Springer, 2016.

[26] Hao Liu, Jiwen Lu, Jianjiang Feng, et Jie Zhou. Learning deep sharable and structural detectors for face alignment. *IEEE Transactions on Image Processing*, 26(4) :1666–1678, 2017.

[27] Heng Yang, Xuming He, Xuhui Jia, et Ioannis Patras. Robust face alignment under occlusion via regional predictive power estimation. *IEEE Transactions on Image Processing*, 24(8) :2393–2403, 2015.

[28] Rajeev Ranjan, Vishal M Patel, et Rama Chellappa. Hyperface : A deep multi-task learning framework for face detection, landmark localization, pose estimation, and gender recognition. *arXiv preprint arXiv :1603.01249*, 2016.

[29] Zhanpeng Zhang, Ping Luo, Chen Change Loy, et Xiaoou Tang. Learning deep representation for face alignment with auxiliary attributes. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 38(5) :918–930, 2016.

[30] Akshay Asthana, Stefanos Zafeiriou, Shiyang Cheng, et Maja Pantic. Robust discriminative response map fitting with constrained local models. Dans *CVPR*, pages 3444–3451, 2013.

[31] Jie Zhang, Shiguang Shan, Meina Kan, et Xilin Chen. Coarse-to-fine auto-encoder networks (cfan) for real-time face alignment. Dans *ECCV*, pages 1–16. Springer, 2014.

[32] Chengchao Qu, Hua Gao, Eduardo Monari, Jurgen Beyerer, et Jean-Philippe Thiran. Towards robust cascaded regression for face alignment in the wild. Dans *CVPR Workshops*, pages 1–9, 2015.

[33] Stefanos Zafeiriou, Cha Zhang, et Zhengyou Zhang. A survey on face detection in the wild : past, present and future. *Computer Vision and Image Understanding*, 138 :1–24, 2015.

[34] Heng Yang, Xuhui Jia, Chen Change Loy, et Peter Robinson. An empirical study of recent face alignment methods. *arXiv preprint arXiv :1511.05049*, 2015.

[35] Jing Yang, Jiankang Deng, Kaihua Zhang, et Qingshan Liu. Facial shape tracking via spatio-temporal cascade shape regression. Dans *ICCV Workshops*, pages 41–49, 2015.

[36] Matej Kristan, Aleš Leonardis, Jiři Matas, Michael Felsberg, Roman Pflugfelder, Luka Čehovin, Tomas Vojir, Gustav Häger, Alan Lukežič, et Gustavo Fernandez. *The Visual Object Tracking VOT2016 Challenge Results*, pages 777–823. Springer International Publishing, Cham, 2016.

[37] Grigorios G. Chrysos, Epameinondas Antonakos, Patrick Snape, Akshay Asthana, et Stefanos Zafeiriou. A comprehensive performance evaluation of deformable face tracking "in-the-wild". *International Journal of Computer Vision*, pages 1–35, 2017.

[38] Enrique Sánchez-Lozano, Brais Martinez, Georgios Tzimiropoulos, et Michel Valstar. Cascaded continuous regression for real-time incremental face tracking. Dans *ECCV*, pages 645–661. Springer, 2016.

[39] Ghassan Hamarneh et Tomas Gustavsson. Deformable spatio-temporal shape models : Extending asm to 2d+ time. Dans *BMVC*, pages 1–10, 2001.

[40] Michal Uricár, Vojtech Franc, et Václav Hlavác. Facial landmark tracking by tree-based deformable part model based detector. Dans *ICCV Workshops*, pages 10–17, 2015.

[41] Shengtao Xiao, Shuicheng Yan, et Ashraf A Kassim. Facial landmark detection via progressive initialization. Dans *ICCV Workshops*, pages 33–40, 2015.

[42] Georgia Rajamanoharan et Timothy F Cootes. Multi-view constrained local models for large head angle facial tracking. Dans *ICCV Workshops*, pages 18–25, 2015.

[43] Yue Wu et Qiang Ji. Shape augmented regression method for face alignment. Dans *ICCV Workshops*, pages 26–32, 2015.

[44] Xi Peng, Rogerio S Feris, Xiaoyu Wang, et Dimitris N Metaxas. A recurrent encoder-decoder network for sequential face alignment. Dans *ECCV*, pages 38–56. Springer, 2016.

[45] Jinwei Gu Xiaodong Yang Shalini De et Mello Jan Kautz. Dynamic facial analysis : From bayesian filtering to recurrent neural network.

[46] Behzad Hasani et Mohammad H Mahoor. Facial expression recognition using enhanced deep 3d convolutional neural networks. *arXiv preprint arXiv :1705.07871*, 2017.