Face alignment: Improving the accuracy of fast models using domain-specific unlabeled data and a teacher-student scheme

Constantino Álvarez Casado and Miguel Bordallo López

Face alignment is a crucial step in multiple face analysis and recognition tasks. The current state-of-the-art is comprised by very slow methods based on deep learning that require computationally heavy inference and very fast methods based on cascades of regressors that lack the ability to cope with complicated cases or extreme poses. We show how, collecting a small subset of unlabeled domain-specific data, we can improve the accuracy of fast-inference models utilizing data annotated by a slower one and a teacher-student architecture. In the proposed solution, we annotate a small subset of facial images belonging to two challenging domains using a slow but more accurate model, and this data is used to incrementally train a fast one. Our results show that by adding as little as a 5% of challenging data, we can reduce the error rate in a specific domain up to 30% without losing any generalization abilities. This training scheme has applicability in numerous computer vision and engineering problems where computational power and model size are constrained by the application and platform or real-time operation is a requirement.

Introduction: Face alignment refers to the method for localizing fiducial facial points on an image. It is a crucial step in multiple face analysis and recognition tasks, with various applications that range from biometrics (identification [1], antispoofing [2] or kinship verification [3]), affective computing (emotion [4] and expression recognition [5]) and even healthcare (medical diagnosis and screening [6]). Even if this challenging problem has been addressed with numerous techniques, it is still far from being solved. Numerous shortcomings and limitations appear especially when real-time performance is required [7].

The current state of the art is comprised by two classes of methods [8]. Slow methods, usually based on deep learning, offer the best accuracy especially the facial images have been taken with extreme poses or in challenging conditions [9][10]. Real-time methods, usually based on cascades of regressors, lack the ability to cope with very complicated cases, but on the other hand they are able to be executed in a few milliseconds, even in the most constrained devices [11][12][13].

In this paper, we propose a training scheme based on a *teacher-student* architecture, that utilizes models created with slow methods and a small subset of unlabeled data to dramatically improve the performance of fast models. The proposed solution annotates a small subset of facial images belonging to challenging domains utilizing a very slow but accurate model, and uses this data to train the fast model. We show that a fast method can improve its performance in domain-specific challenging images without losing any generalization abilities, by adding as little as a 5% of relevant data in the training stage. We illustrate how this technique generalizes to various challenging domains and several face alignment methodologies, demonstrating that the training scheme has applicability in numerous computer vision and engineering problems, where computational power and model size are constrained either by the application and platform or by the real-time requirements of the application.

Experimental setup: In order to evaluate our method, we focus on three face alignment methodologies that are considered to be state-of-the-art for different applications. The first model is trained using a Deep Alignment Network (DAN), one of the methods that has proven to have the best accuracy, at the expense of a relatively big computational time. The second model is based on a Cascade of Regressions of Local Binary Functions (LBF). To provide a comparison and to show the validity of our method for different algorithms and models, we also train a complementary model based on an Ensemble of Regression Trees (ERT).

In this evaluation, we utilize our own C++ single thread implementations of the algorithms which are loosely based on [10] (DAN), [11] (LBF) and [12] (ERT). The results obtained by our implementations closely match the ones published by the original authors of the methods. To measure the computational time, we test the models in two different processors, an Intel(R) Core(TM) i7-6700HQ desktop CPU running at 2.60GHz and a Freescale i.MX6Q Sabrelite ARMv7 Cortex A-9 mobile processor running at 792MHz.

Base models: As the starting point, we train three base models using exactly the same data sets and data augmentation techniques, closely following the training procedures introduced in the original publications. In this context, the base models are trained according to the well-established evaluation approach defined on the 300W competition [14]. In this evaluation protocol, the training data is a subset comprised by the AFW dataset and the training subsets of HELEN and LFPW, resulting in a total of 3148 images. The base training data is augmented first by multiplying five-fold the number of samples while introducing random rotations between -20 and 20 degrees. In addition, the resulting training set is duplicated by mirroring the images for a total of 31480 instances.

The test data proposed by the 300W protocol consists on 2 subsets characterized for their level of difficulty. The 300W common subset consists on 554 images obtained from the test sets of HELEN and FLPW. The 300W challenging subset is comprised by the 135 samples on the IBUG dataset. The full set is the aggregation of both (689 images).

Each one of the images of the are manually annotated with 68 landmarks. The average error of these manual labels is presumed to be from 7 to 8% [15]. Although some of the images include a bounding box generated by a face detector, for fair comparison, we generate our own bounding box both for training and testing, using a simple Viola-Jones detector as implemented by OpenCV.

To measure the error, we utilize the well accepted metric based on the mean distance between the localized landmarks and the ground truth landmarks, all divided by the interpupilary distance (distance between the eye centers), and normalized to a value between 0 and 100.

In addition, in order to characterize the type of errors of each model, we calculate the *"failure rate"*, as the percentage of images that have an error that is larger than an 8%, since errors smaller than that could be considered as annotation errors on the ground truth.

Domain specific challenging datasets: To test the validity of our training approach, we have constructed two domain specific test sets. The Domain-Pitch test set contains 100 images of faces with extreme poses in terms of pitch (head looking up or down). The Domain-Lowlight test set contains 100 images of faces captured in very low illumination or against powerful backlights, presenting poor contrast and dynamic range. Both test sets have been manually annotated with 68 landmarks. A particularity of these two test sets, is that they have been selected so the failure rate of the fast base model (LBF-bas) is a 100% (i.e. all the images present an error larger than 8%). At the same time, the failure rate of the test sets for the slow model (DAN-bas) is 0% (i.e. all the images have an error below 8%). To complement these test sets, we have collected two unlabeled training sets comprised of 150/300/450/600 images without annotation (roughly from 5 to 20% of the original training set) that loosely belong to the same application domain (extreme pitch pose and low illumination). We use these two un-annotated sets to improve the fast models. Example images and their predicted landmarks for the *teacher* model and the *student* model (both base and improved) can be seen in Figure 1.

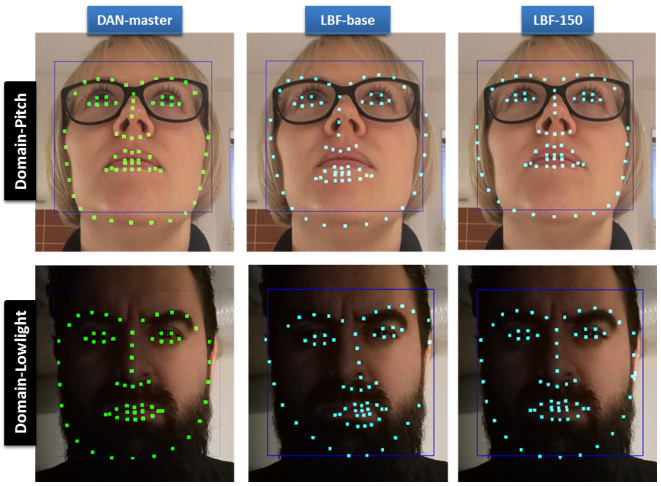

Fig. 1 *Examples of images from the domain specific testing data sets as aligned by the teacher and student models, including success and failure instances.*

Training of the improved models: To improve the accuracy of our fast models, we utilize a *teacher-student* scheme, where the *teacher* is the slow model described before (DAN-mas, based in deep learning, more accurate, but slow). The *teacher* model is is used to annotate 150, 300, 450 and 600 training images of the domain-specific sets. These data, and the landmarks suggested by the *teacher* model is then utilized to incrementally train the *student* models, (LBF-bas, ERT-bas, based on cascades of regressors and very fast). This training results in two sets of improved models (LBF-150 to 600, ERT-150 to 600) that are then evaluated in the testing sets. The improved models, trained with the annotations provided by the *teacher* model, are then able to better predict the landmarks of unseen images belonging to a testing set of the specific domain, without decreasing their performance in the previously evaluated general data-sets.

Results and discussion: The comparative analysis of the experiments with the base and improved *student* models can be seen in Table 1. It can be seen that the teacher model, based on DAN, has execution times that are not suitable for real-time computation when using a CPU. Both the base and improved models are from 100 to 300 times faster on desktop and mobile CPUs. As expected, the inclusion of the proposed training data does not have an impact on computation time, since the models are not fundamentally changed and only the weights of the model vary with the new data.

Table 1: Computation time, average error (%) and failed images (%) on the domain-specific testing subsets.

	comp. time (ms)		Domain Pitch		Domain Lowlight	
Method	CPU	ARM	avg. er.	fail.rate	avg. er.	fail.rate
DAN-tea	300	1000	3.89	0%	4.01	0%
LBF-bas	1-3	3-8	25.5	100%	19.7	100%
LBF-150	1-3	3-8	18.2	76%	15.4	70%
LBF-300	1-3	3-8	16.5	67%	14.8	66%
LBF-450	1-3	3-8	15.4	60%	14.2	63%
LBF-600	1-3	3-8	13.8	52%	13.6	54%
ERT-bas	3-6	3-8	24.2	100%	19.8	100%
ERT-150	3-6	3-8	18.1	84%	17.9	77%
ERT-300	3-6	3-8	16.4	81%	17.5	72%
ERT-450	3-6	3-8	13.7	67%	16.7	67%
ERT-600	3-6	3-8	13.4	56%	16.0	59%

Table 2: Average error (%) and failed images (%) on the general subsets (300W).

	300W-common		300W-challenging		300W-fullset	
Method	avg. er.	fail.rate	avg. er.	fail.rate	avg. er.	fail.rate
DAN-tea	4.4	0.2%	7.57	5.9%	5.03	1.16%
LBF-bas	5.6	9.9%	18.2	82.2%	8.2	23.0%
LBF-150	5.5	8.3%	19.4	77.0%	8.2	23.0%
LBF-300	5.6	10.0%	19.0	77.7%	8.3	23.5%
LBF-450	5.6	9.7%	17.6	79.2%	8.2	23.8%
LBF-600	5.7	10.1%	18.2	76.2%	8.4	23.5%
ERT-bas	5.7	10.1%	17.3	75.5%	8.0	22.9%
ERT-150	5.7	10.4%	16.9	74.0%	7.9	23.1%
ERT-300	5.8	10.7%	16.7	73.3%	8.0	23.3%
ERT-450	5.8	10.8%	17.0	74.8%	8.0	23.6%
ERT-600	5.8	10.4%	17.5	76.2%	8.2	23.3%

When observing the error of the student model (LBF) on both Domain-Pitch and Domain-Lowlight testing sets, it can be seen that the average error is reduced from 15 to 30% adding as little as a 5% of data (150 images). This is a very noticeable reduction that only needs several unlabeled images representative of the challenging problem that we are trying to mitigate. When observing the failure rate, the effect is even more considerable. The improved model is able to accurately label up to 25% of the images where previously was failing.

If we increase the amount of unlabeled data fed to our system progressively from 5% to 20% (600 images), we can see that both error and failure rates can be improved further, reaching a 40% error reduction and reducing the failure rate to 50% for the most favorable cases.

An important characteristic of the training is that the proposed scheme can be utilized without any loss of generalization abilities for the more typical data, as it can be seen in Table 2. When comparing the error and failure rate of both the base and improved models in the 300W test sets, it can be seen that they do not raise noticeably even when a 20% of new data is including in the training stages. For some cases, it can be seen that the error rate indeed improves lightly. This is probably due to the similarity of a few challenging images of the 300W test sets with the ones utilized in our domain specific sets.

Finally, when comparing two different models such as the one based on LBF and the one based on ERT, we can see that the results obtained are comparable, demonstrating that this *teacher-student* architecture for training has applicability in different fast models, independently on the method that is utilized in their regressors.

Conclusion: We have shown that it is possible to improve the performance of very fast face alignment models by just identifying a challenging domain where they do not perform properly, and collecting a small subset of unlabeled representative data. More accurate and slow models can then be used to annotate the set and faster models can be re-trained to improve their accuracy in the selected domain while maintaining its computation times and generalization abilities. Further work includes the study of the combination of data from several challenging domains, the proper characterization of the training data, which could improve the results further, and the identification of the limits of improvement with the addition of more data.

C. Álvarez Casado (*Visidon OY, Finland*) M. Bordallo López (*Center for Machine Vision and Signal Analysis, University of Oulu, Finland*)

E-mail: miguel.bordallo@oulu.fi

References

1 Ding, C., Tao, D.: 'A comprehensive survey on pose-invariant face recognition', *ACM Transactions on intelligent systems and technology*, 2016, , (3), pp. 37

2 Komulainen, J., Boulkenafet, Z., Akhtar, Z.
'Review of face presentation attack detection competitions'.
In: Handbook of Biometric Anti-Spoofing. (Springer, 2019. p. 291

3 Bordallo.Lopez, M., Hadid, A., Boutellaa, E., Goncalves, J., Kostakos, V., Hosio, S.: 'Kinship verification from facial images and videos: human versus. machine', *Machine Vision and Applications*, 2018, **29**, (5), pp. 873–890

4 Sariyanidi, E., Gunes, H., Cavallaro, A.: 'Automatic analysis of facial affect: A survey of registration, representation, and recognition', *IEEE transactions on pattern analysis and machine intelligence*, 2015, **37**, (6), pp. 1113–1133

5 Martinez, B., Valstar, M.F., Jiang, B., Pantic, M.: 'Automatic analysis of facial actions: A survey', *IEEE Transactions on Affective Computing*, 2017,

6 Thevenot, J., Bordallo.López, M., Hadid, A.: 'A survey on computer vision for assistive medical diagnosis from faces', *IEEE Journal of Biomedical and Health Informatics*, 2017, pp. 1–14

7 Wang, N., Gao, X., Tao, D., Yang, H., Li, X.: 'Facial feature point detection: A comprehensive survey', *Neurocomputing*, 2018, **275**, pp. 50–65

8 Jin, X., Tan, X.: 'Face alignment in-the-wild: A survey', *Computer Vision and Image Understanding*, 2017, **162**, pp. 1–22

9 Shi, B., Bai, X., Liu, W., Wang, J.: 'Face alignment with deep regression', *IEEE transactions on neural networks and learning systems*, 2018,

10 Kowalski, M., Naruniec, J., Trzcinski, T.
'Deep alignment network: A convolutional neural network for robust face alignment'.
In: Proceedings of the International Conference on Computer Vision & Pattern Recognition (CVPRW), Faces-in-the-wild Workshop/Challenge. vol. 3. (, 2017. p. 6

11 Ren, S., Cao, X., Wei, Y., Sun, J.
'Face alignment at 3000 fps via regressing local binary features'.
In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. (, 2014. pp. 1685–1692

12 Kazemi, V., Sullivan, J.
'One millisecond face alignment with an ensemble of regression trees'.
In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. (, 2014. pp. 1867–1874

13 Ren, S., Cao, X., Wei, Y., Sun, J.: 'Face alignment via regressing local binary features', *IEEE Transactions on Image Processing*, 2016, **25**, (3), pp. 1233–1245

14 Sagonas, C., Antonakos, E., Tzimiropoulos, G., Zafeiriou, S., Pantic, M.: '300 faces in-the-wild challenge: Database and results', *Image and vision computing*, 2016, **47**, pp. 3–18

15 Sagonas, C., Tzimiropoulos, G., Zafeiriou, S., Pantic, M.
'A semi-automatic methodology for facial landmark annotation'.
In: Proceedings of the IEEE conference on computer vision and pattern recognition workshops. (, 2013. pp. 896–903