Journal of Visual Communication and Image Representation 55 (2018) 841–852

Contents lists available at ScienceDirect

Journal of Visual Communication and Image Representation

journal homepage: www.elsevier.com/locate/jvci

Sampling strategies for performance improvement in cascaded face regression[☆]

Romuald Perrot[*], Pascal Bourdon, David Helbert

XLIM-ASALI University of Poitiers, UMR CNRS, 7252, France

ARTICLE INFO

Keywords:
Face landmarking
Regression
Sampling
Data augmentation

ABSTRACT

Automatic face landmarking has received a lot of attention in the past decades. It is now mature enough to be implemented in fully autonomous video systems. As cascade-of-regression based algorithms have become state of the art in such systems, two major (and still relevant) sources of interest have slowly faded away: the need for semantic-driven learning beyond ground truth annotation, and full video chain performance *i.e.* tracking efficiency, which in the case of said methods strongly relates to their robustness towards shape initialization before fitting. In this paper, we investigate how data sampling using face priors can affect their performance in terms of convergence and robustness. We propose new strategies based on said priors to overcome inconsistencies observed during cascade-of-regression learning on purely random sampling-based stages. We will show that simple choices can be easily integrated within regression-based face tracking systems to increase accuracy and robustness.

1. Introduction

Face tracking or face landmarking is an active topic and also a key tool for image and video analysis: authentification, emotion detection or face transfer are some of the applications relying on face landmarking.

As of today, cascade-of-regression is still one of the most popular methods employed for face landmarking when low computational resources are required [1–3]. While deep architectures with convolution structure (*i.e.* Convolutional Neural Networks) have been found highly effective for this task very recently [4–6], their strong reliance on parallel computation efficiency and intensive use of Graphics Processing Unit (GPU) ressources prevent them from being used on low-cost, low-power embedded systems. It is important to stress out that our most of our research work on face analysis is indeed aimed at such systems.

Like deep neural networks, cascade-of-regression algorithms rely on supervised learning: starting from a learning database, a regression model is built then used later for face fitting. The database may not be exhaustive, thus it requires an extra step called *data augmentation* to increase variation of the input. To our concern, authors do not explore sampling strategies and tend to rely on blind uniform sampling, leading to good fitting performance on a single, pre-initialized frame basis but to very poor full video chain processing choices. Indeed, cascade-of-regression alignment has the well-known drawback that it is essentially aimed at static images by design and rely on accurate initializations from potentially computationally expensive face detectors [7]. Beyond face detection performance, the influence of inaccurate shape initialization is rarely investigated, at least using realistic scenarios and not just easily removable zero-centered Gaussian noise, leading to either tedious frame-by-frame detection tasks or unstable bounding box tracking strategies.

The model used in cascaded face landmarking is a tree built with a random process. Nodes of the tree are created upon a set of features, which is inherently a limited set. As a result, a sampling scheme should again be employed to define this set. Like data augmentation, authors only rely on uniform sampling and the resulting trained models may not be optimal, with final step decisions leading to regression directions cancelling one another out, as we will illustrate later on.

Our contributions in this paper concern the study of sampling strategies used during two steps of the regression learning scheme: data augmentation and feature sampling. For both steps, we investigate several sampling methods found in literature as well as new propositions and investigate their impact on face fitting quality. Readers should keep in mind that while our field of investigation is primarily cascade-of-regression alignment, most other methods can or already benefit from such sampling strategies. In details, we propose:

1. Four sampling schemes for data augmentation, taking into account common knowledge about face geometry and dynamics;

[☆] This paper has been recommended for acceptance by Jiwen Lu.

[*] Corresponding author.

E-mail addresses: romuald.perrot@univ-poitiers.fr (R. Perrot), pascal.bourdon@univ-poitiers.fr (P. Bourdon), david.helbert@univ-poitiers.fr (D. Helbert).

https://doi.org/10.1016/j.jvcir.2018.07.006
Received 10 November 2017; Received in revised form 8 June 2018; Accepted 7 July 2018
Available online 01 September 2018

2. Two sampling schemes for features generation, with two opposite directions: better space coverage and landmark importance;
3. An analysis of semantic-driven sampling strategies compared to conventional blind sampling.

The paper is structured as follows. Section 2 discusses previous work on feature sampling and data augmentation for face landmarking. Section 3 presents regression-based methods. In Section 4 we expose various sampling schemes for building augmented sets of groundtruth shapes. Section 5 details sampling schemes for building feature sets. In Section 6 we study the results of each sampling scheme in the context of face landmarking, and set a performance benchmark of our algorithm using a challenging faces-in-the-wild dataset, namely the 300W competition test-set [8,9]. Finally Section 7 concludes the paper.

2. Related work

In this section we only reference papers that are fundamental to understand our work. A full survey of face landmarking is beyond the scope of this paper. Readers can refer to recent surveys such as [10] or [11] for a detailed overview.

2.1. Face alignment

Historically, authors categorize face landmarking into three main methods: Active Shape and Active Appearance Models (ASM/AAM) [12,13], Constrained Local Models (CLM) [14] and regression models [15]. AAM conjointly learn global texture and shape models, with face fitting consisting in minimizing the difference between a target face image and a deformable parametric texture model. Unfortunately, the idea of a global texture model is a major issue since it tends to drive the fitting process as a whole, turning common occurrences such as occlusion or light changes into a threat to result quality. To increase robustness against occlusions, CLM methods employ a local texture model around each landmark. While some CLM implementations allow interactive frame rate [16], they are usually computationally expensive and require high-performance hardware.

Regression-based methods have been claimed to enable both high-performance and high-robustness in face landmarking, even on limited hardware such as smart devices. Dollar et al. [15] introduce the cascaded pose regression method, where a shape is progressively refined to a target shape. Cao et al. [1] enhance regression using a new shape indexation scheme and a boosted two-level cascade of regression. Kazemi et al. [2] provide a high-performance regression system with a simplified initialization stage and gradient boosted cascade building. At the same time, Ren et al. [3] announce three times speed-up using customized local binary features. While impressive performances are reported, almost every method suffers from the same issues. Yang et al. [17] show that such methods are highly sensitive to prior face detection performance, said detection being a mandatory step for initialization. As a result, despite solution proposals such as combined detection/regression [18], the issue of robustness towards initialization is still considered an open issue.

2.2. About deep architectures

Very recently, deep neural networks have been applied to face alignment, either for 2D landmarking [5,19] or 3D landmarking [4,6]. Such methods rely on training cascades of Convolutional Neural Networks (CNNs) to compute features around landmarks using all pixels of the images as input. While these methods have shown impressive results regarding fitting precision, especially in videos [20], they all rely on a computationally intensive process that is often incompatible with real-time fitting or even just computation on low-resource, low-power embedded hardware. Resource and time–cost effectiveness is considered a rationale for cascades of regression, as defended by Kazemi et al.'s 1000fps [2] or Ren et al.'s 3000fps [3] experiments. In [5] Trigeorgis et al. criticize the fact that the binary/tree-based features commonly used in cascade-of-regression methods, being just simple pixel intensity differences, cannot be learnt in an end-to-end manner like, as opposed to convolutional features. While the above remark is reliable and well illustrated by the authors, the use of (well-named) weak learners such as random ferns [1,21] or random forests [2,3] within a boosting framework is precisely the reason why such methods require so little computational ressources, as opposed to convolutional masks. Indeed, we consider our introduction of face priors in a data-driven manner during the sampling phase (*importance sampling*) as a step towards semantic training which echoes Trigeorgis et al./Mnemonic Descent Method (MDM)'s natural learning of head pose partitions [5] or Cao's coarse-to-fine hierarchy obtained with shape-constrained regression [1]. Comparative results between our algorithm and current state-of-art face landmarking methods (including MDM) will be provided in Section 6 to demonstrate its relevance in terms of precision and computational costs. Moreover, we will provide an in-depth analysis of regression behaviour between random sampling and importance sampling, illustrated by Figs. 16 and 17, showing how importance sampling can prevent regression directions from *cancelling one another out* despite the fact that regressors are learnt independently, which was another source of criticism pointed by Trigeorgis et al.

2.3. Features sampling

Most studies on features used in face landmarking have been done with robustness against face transformations in mind (mainly rotation and perspective transformation). Cao et al. [1] use shape-indexation, Burgos et al. [21] introduce interpolated shape features, which is later enhanced by Cao [22] using barycentric coordinates. Unfortunately, feature pool selection has not been well investigated. Dollar et al. [15], use uniform sampling; reference methods [1,21,2,3] are based on Dollar's work thus use the same sampling strategy. To our concern, the only reference method that employs another method is the work of Cao et al. [22] where a Gaussian distribution over the unit square is used, although no specific justification or comparison with uniform sampling is provided by the authors. Kazemi et al. [2] observed that feature selection using distance priors leads to better fitting performance but feature generation is still based on uniform sampling.

2.4. Data augmentation sampling

Some authors have studied the impact of data augmentation on classification performance [23]. As an example, Krizhevsky et al. [24] perform Principal Component Analysis (PCA) on Red-Green–Blue (RGB) pixel values in a deep learning architecture to achieve the best results (at the time of publication) on the famous ImageNet classification challenge. De Vries et al. [25] also used data space to perform data augmentation. To our knowledge, such complex data-space augmentation methods have not been applied to regression-based face landmarking, and only blind, uniform selection of shapes is used during data augmentation. As an example, as PCA modelling has been criticized as the cause of ASM/AAM/CLM's failure to fit in-the-wild face shapes, anything PCA-related has been seemingly discarded from cascade-of-regression landmarking, including their use for data augmentation documented in [13,26].

It is interesting to note that in the context of deep learning, where the number of training samples is often much higher, some works [27,28] have been conducted in the opposite direction: sampling the training set to generate a smaller set that leads to the same fitting/classification error. The main aim is to reduce computational training cost.

In this paper, we propose new strategies for both data augmentation and feature sampling, where face semantic integration is induced implicitly by priors regarding data representation models and space