Towards Spatio-Temporal Face Alignment in Unconstrained Conditions
Romain Belmonte, Nacim Ihaddadene, Pierre Tirilly, Chaabane Djeraba, Ioan Marius Bilasco

To cite this version:
Romain Belmonte, Nacim Ihaddadene, Pierre Tirilly, Chaabane Djeraba, Ioan Marius Bilasco. Towards Spatio-Temporal Face Alignment in Unconstrained Conditions. VISAPP, Jan 2018, Funchal, Portugal. hal-01644813

HAL Id: hal-01644813
https://hal.science/hal-01644813
Submitted on 29 Sep 2018

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L’archive ouverte pluridisciplinaire HAL, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d’enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.
Towards Spatio-Temporal Face Alignment in Unconstrained Conditions

R. Belmonte\textsuperscript{1,3}, N. Ihaddadene\textsuperscript{1}, P. Tirilly\textsuperscript{2}, M. Bilasco\textsuperscript{3} and C. Djeraba\textsuperscript{2}

\textsuperscript{1}ISEN Lille, Yncrea Hauts-de-France, France
\textsuperscript{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
\textsuperscript{3}Univ. Lille, CNRS, Centrale Lille, UMR 9189 - CRIStAL - Centre de Recherche en Informatique Signal et Automatique de Lille, F-59000 Lille, France

\{romain.belmonte, nacim.Ihaddadene\}@yncrea.fr, \{pierre.tirilly, marius.bilasco, chaabane.djeraba\}@univ-lille1.fr

Keywords: Face Alignment, Spatio-Temporal Models, Landmark Localization, Video-based, Unconstrained Conditions

Abstract: Face alignment is an essential task for many applications. Its objective is to locate feature points on the face, in order to identify its geometric structure. Under unconstrained conditions, the different variations that may occur in the visual context, together with the instability of face detection, make it a difficult problem to solve. While many methods have been proposed, their performances under these constraints are still not satisfactory. In this article, we claim that face alignment should be studied using image sequences rather than still images, as it has been done so far. We show the importance of taking into consideration the temporal information under unconstrained conditions.

1 INTRODUCTION

The problem of face alignment, also called facial landmark localization, has raised much interest and experienced rapid progress in recent years (Jin and Tan, 2016). Given the position and size of a face, the alignment process, illustrated in Figure 1, consists in determining the geometry of the face components containing the most useful information (e.g., eyes, nose, mouth). This ability to model non-rigid facial structures is now used in various fields such as face analysis (e.g., identification, expression recognition) (Sun et al., 2014), human-computer interaction (Akakin and Sankur, 2009) or information retrieval (Park and Jain, 2010). However, despite the large number of methods available in the literature, the performance of face alignment under unconstrained conditions remains limited (Sagonas et al., 2016).

Even today, this problem continues to be studied using still images (Jin and Tan, 2016). Yet, due to the ubiquity of video sensors, the vast majority of applications rely on image sequences. In addition, many tasks related to face analysis or, more broadly, human behavior analysis have leveraged temporal information (Simonyan and Zisserman, 2014; Fan et al., 2016). The first survey on face alignment have already suggested to study the problem using image sequences, but without providing actual arguments (Çeliktutan et al., 2013). Our motivation in this paper is to show that taking temporal information into account for this problem could greatly contribute to improve performance under unconstrained conditions.

This article is structured as follows: in Section 2, we describe the reasons that led us towards spatio-temporal approaches for face alignment. In particular, we review the existing solutions and put into perspective the performance of the most recent methods. Section 3 shows the benefits of temporal information under unconstrained conditions. Finally, we conclude with Section 4.

Figure 1: Face alignment process: (a) original image, (b) face detection, (c) face alignment. Images from Menpo (Zafeiriou et al., 2017).
Table 1: Datasets captured under unconstrained conditions.

2 LITERATURE REVIEW

In this section, the datasets and evaluation metrics used for face alignment under unconstrained conditions are discussed. The main categories of methods from literature are reviewed and their performance is analyzed.

2.1 Datasets and Metrics

In recent years, many datasets for face alignment have been made available to the scientific community (cf. Table 1). The images included in these datasets are collected on social networks or image search services such as Google, Flickr, or Facebook, bringing more realism to the data. The annotation is performed either manually or semi-automatically, sometimes with the help of the Amazon Mechanical Turk platform. The quality of the annotations, however, may vary (Sagonas et al., 2016; Bulat and Tzimiropoulos, 2017; Çeliktutan et al., 2013). The annotation scheme used (i.e., position and number of landmarks) may also differ from one dataset to another. Currently, the scheme composed of 68 landmarks (Gross et al., 2010), illustrated in Figure 1, is the most widely used. Since this scheme does not fit at extreme poses, a 39-point-based scheme has recently been proposed for profile faces (Zafeiriou et al., 2017) (see Figure 1).

Today, due in particular to the emergence of deep learning techniques, it may be necessary to have a large number of annotated images, which existing datasets do not necessarily provide. In the literature, various augmentation methods are used to circumvent this problem. Some operations (e.g., rotation, mirroring, disturbance of the detection window position and size) can be applied to images to generate new training samples. Other more complex processes such as the generation of synthetic images may also be used (Zhu et al., 2016).

Moreover, it is crucial to have data that is representative of the problem in order to answer it. Static datasets do not cover all the difficulties encountered by applications. The constraints related to the movements of persons or cameras are not currently considered. A dataset composed of image sequences captured under unconstrained conditions has recently been published by (Shen et al., 2015) (cf. Figure 2). This data, in addition to being more realistic, provides some clues in favor of the use of temporal information for face alignment.

To evaluate predictions on this data, the mean square error normalized by the interocular distance (NMSE) is generally used. Beyond an error value of 8%, landmarks are mostly not located correctly and the prediction is considered as a failure. Normalization by the interocular distance, although not very robust to extreme poses, is the most widespread. Other normalization factors are sometimes used, such as the diagonal of the detection window. On a set of images, average NMSE is the simplest and most intuitive evaluation metric. However, it can be strongly affected by a few outliers. A graphical representation of the error distribution function is therefore increasingly used. It corresponds to the proportion of images for which the error is less than or equal to a certain threshold (e.g.,
### Table 2: Performances of recent literature methods. Normalized mean squared error is reported.

<table>
<thead>
<tr>
<th>Method</th>
<th>300W-A</th>
<th>300W-B</th>
<th>300W</th>
</tr>
</thead>
<tbody>
<tr>
<td>RCPR (Burgos-Artizzi et al., 2013)</td>
<td>6.18</td>
<td>17.26</td>
<td>8.35</td>
</tr>
<tr>
<td>SDM (Xiong and De la Torre, 2013)</td>
<td>5.57</td>
<td>15.40</td>
<td>7.50</td>
</tr>
<tr>
<td>ESR (Cao et al., 2014)</td>
<td>5.28</td>
<td>17.00</td>
<td>7.58</td>
</tr>
<tr>
<td>CFAN (Zhang et al., 2014)</td>
<td>5.50</td>
<td>16.78</td>
<td>7.69</td>
</tr>
<tr>
<td>LBF (Ren et al., 2014)</td>
<td>4.95</td>
<td>11.98</td>
<td>6.32</td>
</tr>
<tr>
<td>CFSS (Zhu et al., 2015)</td>
<td>4.73</td>
<td>9.98</td>
<td>5.76</td>
</tr>
<tr>
<td>RPPE (Yang et al., 2015a)</td>
<td>5.50</td>
<td>11.57</td>
<td>6.69</td>
</tr>
<tr>
<td>3DDEA (Zhu et al., 2016)</td>
<td>6.15</td>
<td>10.59</td>
<td>7.01</td>
</tr>
<tr>
<td>TCDCN (Zhang et al., 2016)</td>
<td>4.80</td>
<td>8.60</td>
<td>5.54</td>
</tr>
<tr>
<td>RAR (Xiao et al., 2016)</td>
<td>4.12</td>
<td><strong>8.35</strong></td>
<td><strong>4.94</strong></td>
</tr>
<tr>
<td>RCFA (Wang et al., 2016)</td>
<td>4.03</td>
<td>9.85</td>
<td>5.32</td>
</tr>
<tr>
<td>R-DSSD (Liu et al., 2017a)</td>
<td>4.16</td>
<td>9.20</td>
<td>5.59</td>
</tr>
</tbody>
</table>

Despite the quantity of methods proposed in the literature and the recent major advances, we can see from these results that the problems encountered under unconstrained conditions are still far from being solved. Because of their significant influence on facial appearance, variations in pose and occlusions are among the most difficult challenges (Shen et al., 2015; Sagonas et al., 2016). We show in Section 3 how temporal approaches could help to address these problems.
<table>
<thead>
<tr>
<th>Method</th>
<th>Category 1</th>
<th>Category 2</th>
<th>Category 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Uricar et al., 2015)</td>
<td>0.657</td>
<td>0.677</td>
<td>0.574</td>
</tr>
<tr>
<td>(Xiao et al., 2015)</td>
<td>0.760</td>
<td>0.782</td>
<td>0.695</td>
</tr>
<tr>
<td>(Rajamanoharan and Cootes, 2015)</td>
<td>0.735</td>
<td>0.717</td>
<td>0.659</td>
</tr>
<tr>
<td>(Wu and Ji, 2015)</td>
<td>0.674</td>
<td>0.732</td>
<td>0.602</td>
</tr>
<tr>
<td>(Zhu et al., 2015; Danelljan et al., 2015)</td>
<td>0.729</td>
<td>0.777</td>
<td>0.684</td>
</tr>
<tr>
<td>(Yang et al., 2015c)</td>
<td>0.791</td>
<td>0.788</td>
<td>0.710</td>
</tr>
</tbody>
</table>

Table 3: Comparison of methods from (Shen et al., 2015; Chrysos et al., 2017), on the 3 categories of 300VW. Area under the curve (AUC) and failure rate (FR) are reported.

3 THE BENEFITS OF TEMPORAL INFORMATION

In this section, we show how temporal information can be beneficial to the problem of face alignment under unconstrained conditions. Issues raised by the dependence to face detection are first discussed. Constraints including the trajectories of the landmarks are then pointed out as solutions to enhance the robustness and quality of face alignment.

3.1 Rigid and Non-rigid Tracking

Face detection under unconstrained conditions is a complex problem to solve (Zafeiriou et al., 2015). Given its role in face alignment, (Yang et al., 2015b) studied the dependence between these two tasks. They showed a high sensitivity of alignment to detection quality. Thus, besides detection failures, factors such as variations in the scale and position of the detection window may disrupt the alignment.

A solution to avoid the dependence on face detection is to perform non-rigid face tracking. (Shen et al., 2015) recently proposed a comparative analysis of current non-rigid face tracking methods. The results are referenced in Table 3. The most popular strategy is tracking by detection, that is, face detection and alignment on each image independently, without making use of adjacent frames. An alternative to tracking by detection is to use a substitute for the detection such as a generic (i.e., rigid) tracking algorithm (Danelljan et al., 2015). One of the advantages of generic tracking algorithms is that they are able to take into account some variations in the appearance of the target object during tracking (Kristan et al., 2016). (Chrysos et al., 2017) evaluate this strategy and compare it to tracking by detection. In general, generic tracking makes it possible to be more robust to the challenges encountered under unconstrained conditions. It is, however, probable that, as with detection, alignment is sensitive to changes in the tracking window.

(Yang et al., 2015c) avoid face detection at each frame or rigid tracking by proposing a spatio-temporal cascade regression. They initialize the shape at the current frame from the similarity parameters at the previous frame. They incorporate a re-initialization mechanism based on the quality of the prediction in order to avoid any drift in the alignment. Their method can greatly reduce the failure rate while improving overall performance (cf. Table 3). (Sánchez-Lozano et al., 2016) propose an incremental cascaded continuous regression. In contrast to (Yang et al., 2015c) which retains a generic model after learning, here a pre-trained model is updated online to become specific to each person during tracking. This type of approach yields better results than a generic model. In the end, non-rigid tracking produces more accurate fitting than tracking by detection. It takes advantage of adjacent frames to improve initialization and variations in appearance to increasingly become person-specific. Yet, other information such as the trajectories of the landmarks through the image sequence seems relevant to consider (Hamarneh and Gustavsson, 2001). This will be discussed in detail in the next subsection.

3.2 Additional Constraints

Whether explicit or implicit, the shape constraints present in most alignment methods are crucial to obtain good performance in unconstrained conditions. In image sequences, an additional constraint may be applied to the trajectories of the landmarks. Bayesian filters such as Kalman filters or particle filters can be used for this purpose. However, (De and Kautz, 2017) highlighted the marginal gain provided by these approaches and showed the superiority of recurrent neural networks for dynamic face analysis.

(Peng et al., 2016; Hou et al., 2017) also use recurrent neural networks to exploit the dynamic features of the face. They compare a recurrent and a non-recurrent version of their network and show that recurrent learning improves the stability of predictions...
and the robustness to occlusions, variations in pose and expressions. The performances of their methods are referenced in Table 4. Taking into account temporal information results in a decrease of the average error of more than 1% compared to state-of-the-art static methods.

<table>
<thead>
<tr>
<th>Approach</th>
<th>Method</th>
<th>300VW</th>
</tr>
</thead>
<tbody>
<tr>
<td>Static</td>
<td>(Zhang et al., 2016)</td>
<td>7.59</td>
</tr>
<tr>
<td>Dynamic</td>
<td>(Zhang et al., 2016)</td>
<td>6.25</td>
</tr>
<tr>
<td></td>
<td>(De and Kautz, 2017)</td>
<td>6.16</td>
</tr>
<tr>
<td></td>
<td>(Liu et al., 2017b)</td>
<td>5.59</td>
</tr>
</tbody>
</table>

Table 4: Comparison of a static multi-task method (Zhang et al., 2016), with three dynamic methods based on recurrent neural networks (Peng et al., 2016; De and Kautz, 2017; Liu et al., 2017b). Normalized mean squared error is reported.

More recently, (Liu et al., 2017b) propose a two-stream recurrent network composed of a spatial stream that preserve the holistic facial shape structure and a temporal stream that discover shape-sensitive and spatio-temporal features. Their method outperforms those based on a single stream (see Table 4). These are only the beginnings of the use of temporal information for the problem of face alignment. Recurrent neural networks, although advantageous, are capable of characterizing only the global motion. In other related tasks, local motion (i.e., over a few frames) sometimes associated with global motion has led to interesting results (Hasani and Mahoor, 2017; Fan et al., 2016) and could be equally beneficial to face alignment.

4 CONCLUSION

In this paper we presented a review of current work on face alignment under unconstrained conditions. This problem has been studied on still images for several decades, despite applications being mainly based on image sequences. Moreover, many tasks related to face analysis or, more broadly, human behavior analysis have leveraged temporal information. To the best of our knowledge, this is one of the first surveys to quantify the difference between static and dynamic face alignment methods. Especially, we have shown that taking into consideration the temporal information greatly contribute to overcome unconstrained conditions challenges. Recent work that exploit the dynamic features of the face are able to improve initialization and stability of predictions leading to more accurate fitting. Nevertheless, there is still much to be done.

REFERENCES


