A Comprehensive Performance Evaluation of Deformable Face Tracking "In-the-Wild"

Grigorios G. Chrysos[1] · Epameinondas Antonakos[1] · Patrick Snape[1] · Akshay Asthana[2] · Stefanos Zafeiriou[1,3]

Received: 15 March 2016 / Accepted: 10 February 2017

Abstract Recently, technologies such as face detection, facial landmark localisation and face recognition and verification have matured enough to provide effective and efficient solutions for imagery captured under arbitrary conditions (referred to as "in-the-wild"). This is partially attributed to the fact that comprehensive "in-the-wild" benchmarks have been developed for face detection, landmark localisation and recognition/verification. A very important technology that has not been thoroughly evaluated yet is deformable face tracking "in-the-wild". Until now, the performance has mainly been assessed qualitatively by visually assessing the result of a deformable face tracking technology on short videos. In this paper, we perform the first, to the best of our knowledge, thorough evaluation of state-of-the-art deformable face tracking pipelines using the recently introduced 300 VW benchmark. We evaluate many different architectures focusing mainly on the task of on-line deformable face tracking. In particular, we compare the following general strategies: (a) generic face detection plus generic facial landmark localisation, (b) generic model free tracking plus generic facial landmark localisation, as well as (c) hybrid approaches using state-of-the-art face detection, model free tracking and facial landmark localisation technologies. Our evaluation reveals future avenues for further research on the topic.

Keywords Deformable face tracking · Face detection · Model free tracking · Facial landmark localisation · Long-term tracking

1 Introduction

The human face is arguably among the most well-studied deformable objects in the field of Computer Vision. This is due to the many roles it has in numerous applications. For example, accurate detection of faces is an essential step for tasks such as controller-free gaming, surveillance, digital photo album organization, image tagging, etc. Additionally, detection of facial features plays a crucial role for facial behaviour analysis, facial attributes analysis (e.g., gender and age recognition, etc.), facial image editing (e.g., digital make-up, etc.), surveillance, sign language recognition, lip reading, human-computer and human-robot interaction. In this work, we study the deformable face tracking task and we develop

Communicated by Lourdes Agapito.

E. Antonakos and P. Snape contributed equally and have joint second authorship.

Electronic supplementary material The online version of this article (doi:10.1007/s11263-017-0999-5) contains supplementary material, which is available to authorized users.

✉ Grigorios G. Chrysos
g.chrysos@imperial.ac.uk

Epameinondas Antonakos
e.antonakos@imperial.ac.uk

Patrick Snape
p.snape@imperial.ac.uk

Akshay Asthana
akshay.asthana@seeingmachines.com

Stefanos Zafeiriou
s.zafeiriou@imperial.ac.uk

[1] Department of Computing, Imperial College London, 180 Queen's Gate, London SW7 2AZ, UK

[2] Seeing Machines Ltd., Level 1, 11 Lonsdale St, Braddon, ACT 2612, Australia

[3] Center for Machine Vision and Signal Analysis, University of Oulu, Oulu, Finland

Published online: 25 February 2017

the first, to the best of our knowledge, comprehensive evaluation of multiple deformable face tracking pipelines.

Current research has been monopolised by the tasks of *face detection*, *facial landmark localisation* and *face recognition or verification*. Firstly, face detection, despite having permeated many forms of modern technology such as digital cameras and social networking, is still a challenging problem and a popular line of research, as shown by the recent surveys of Jain and Learned-Miller (2010), Zhang and Zhang (2010), Zafeiriou et al. (2015). Although face detection on well-lit frontal facial images can be performed reliably on an embedded device, face detection on arbitrary images of people is still extremely challenging (Jain and Learned-Miller 2010). Images of faces under these unconstrained conditions are commonly referred to as "in-the-wild" and may include scenarios such as extreme facial pose, defocus, faces occupying a very small number of pixels or occlusions. Given the fact that face detection is still regarded as a challenging task, many generic object detection architectures such as Yan et al. (2014), King (2015) are either directly assessed on in-the-wild facial data, or are appropriately modified in order to explicitly perform face detection as done by Zhu and Ramanan (2012), Felzenszwalb and Huttenlocher (2005). The interested reader may refer to the most recent survey by Zafeiriou et al. (2015) for more information on in-the-wild face detection. The problem of localising facial landmarks that correspond to fiducial facial parts (e.g., eyes, mouth, etc.) is still extremely challenging and has only been possible to perform reliably relatively recently. Although the history of facial landmark localisation spans back many decades (Cootes et al. 1995, 2001), the ability to accurately recover facial landmarks on in-the-wild images has only become possible in recent years (Matthews and Baker 2004; Papandreou and Maragos 2008; Saragih et al. 2011; Cao et al. 2014). Much of this progress can be attributed to the release of large annotated datasets of facial landmarks (Sagonas et al. 2013b, a; Zhu and Ramanan 2012; Le et al. 2012; Belhumeur et al. 2013; Köstinger et al. 2011) and very recently the area of facial landmark localisation has become extremely competitive with recent works including Xiong and De la Torre (2013), Ren et al. (2014), Kazemi and Sullivan (2014), Zhu et al. (2015), Tzimiropoulos (2015). For a recent evaluation of facial landmark localisation methods the interested reader may refer to the survey by Wang et al. (2014) and to the results of the 300 W competition by Sagonas et al. (2015). Finally, face recognition and verification are extremely popular lines of research. For the past two decades, the majority of statistical machine learning algorithms spanning from linear/non-linear subspace learning techniques (De la Torre 2012; Kokiopoulou et al. 2011) to deep convolutional neural networks (DCNNs) (Taigman et al. 2014; Schroff et al. 2015; Parkhi et al. 2015) have been applied to the problem of face recognition and verification. Recently, due to the revival of DCNNs, as well as the development of graphics processing units (GPUs), remarkable face verification performance has been reported (Taigman et al. 2014). The interested reader may refer to the recent survey by Learned-Miller et al. (2016) as well as the most popular benchmark for face verification in-the-wild in Huang et al. (2007).

In all of the aforementioned fields, significant progress has been reported in recent years. The primary reasons behind these advances are:

- *The collection and annotation of large databases* Given the abundance of facial images available primarily through the Internet via services such as Flickr, Google Images and Facebook, the collection of facial images is extremely simple. Some examples of large databases for face detection are FDDB (Jain and Learned-Miller 2010), AFW (Zhu and Ramanan 2012) and LFW (Huang et al. 2007). Similar large-scale databases for facial landmark localisation include 300 W (Sagonas et al. 2013b) LFPW (Belhumeur et al. 2013), AFLW (Köstinger et al. 2011) and HELEN (Le et al. 2012). Similarly, for face recognition there exists LFW (Huang et al. 2007), FRVT (Phillips et al. 2000) and the recently introduced Janus database (IJB-A) (Klare et al. 2015).
- *The establishment of in-the-wild benchmarks and challenges* that provide a fair comparison between state of the art techniques. FDDB (Jain and Learned-Miller 2010), 300 W (Sagonas et al. 2013a, 2015) and Janus (Klare et al. 2015) are the most characteristic examples for face detection, facial landmark localisation and face recognition, respectively.

Contrary to face detection, facial landmark localisation and face recognition, the problem of *deformable face tracking* across long-term sequences has yet to attract much attention, despite its crucial role in numerous applications. Given the fact that cameras are embedded in many common electronic devices, it is surprising that current research has not yet focused towards providing robust and accurate solutions for long-term deformable tracking. Almost all face-based applications, including facial behaviour analysis, lip reading, surveillance, human-computer and human-robot interaction etc., require accurate *continuous tracking* of the facial landmarks. The facial landmarks are commonly used as input signals of higher-level methodologies to compute motion dynamics and deformations. The performance of currently available technologies for facial deformable tracking has not been properly assessed (Yacoob and Davis 1996; Essa et al. 1996, 1997; Decarlo and Metaxas 2000; Koelstra et al. 2010; Snape et al. 2015). This is attributed to the fact that, until recently, there was no established benchmark for the task. At ICCV 2015, the first benchmark for facial landmark tracking (so-called 300 VW) was presented by Shen et al.

(2015), providing a large number of annotated videos captured in-the-wild.[1] In particular, the benchmark provides 114 videos with average duration around 1 minute, split into three categories of increasing difficulty. The frames of all videos (218595 in total) were annotated by applying semi-automatic procedures, as shown in Chrysos et al. (2015). Five different facial tracking methodologies were evaluated in the benchmark (Rajamanoharan and Cootes 2015; Yang et al. 2015a; Wu and Ji 2015; Uricar and Franc 2015; Xiao et al. 2015) and the results are indicative of the current state-of-the-art performance.

In this paper, we make a significant step further and develop the first, to the best of our knowledge, comprehensive evaluation of multiple deformable face tracking pipelines. In particular, we assess:

- A pipeline which combines a generic face detection algorithm with a facial landmark localisation method. This pipeline is typically assumed in the related tracking papers, e.g. Wolf et al. (2011), Best-Rowden et al. (2013), Chrysos et al. (2015), as well as in various implementations that are (publicly) available, e.g. King (2009), Asthana et al. (2014), Chrysos et al. (2015), and the demos given in various conferences. The pipeline is fairly robust since the probability of drifting is reduced due to the application of the face detector at each frame. Nevertheless, it does not exploit the dynamic characteristics of the tracked face. Several state-of-the-art face detectors as well as facial landmark localisation methodologies are evaluated in this pipeline.
- A pipeline which combines a model free tracking system with a facial landmark localisation method. This approach takes into account the dynamic nature of the tracked face, but is susceptible to drifting and thus losing the tracked object. We evaluate the combinations of multiple state-of-the-art model free trackers, as well as landmark localisation techniques.
- Hybrid pipelines that include mechanisms for detecting tracking failures and performing re-initialisation, as well as using models for ensuring robust tracking.

Some of the above pipelines were used extensively by practitioners, especially the first one. Nevertheless, to the best of our knowledge, this is the first paper that explicitly refers to the various alternatives and provides a thorough examination of the different components of the pipelines (i.e., detectors, trackers, smoothing, landmark localisation etc.).

Summarising, the findings of our evaluation show that current face detection and model free tracking technologies are advanced enough so that even a naive combination with landmark localisation techniques is adequate to achieve state-of-the-art performance on deformable face tracking. Specifically, we experimentally show that model free tracking based pipelines are very accurate when applied on videos with moderate lighting and pose circumstances. Furthermore, the combination of state-of-the-art face detectors with landmark localisation systems demonstrates excellent performance with surprisingly high true positive rate on videos captured under arbitrary conditions (extreme lighting, pose, occlusions, etc.). Moreover, we show that hybrid approaches provide only a marginal improvement, which is not worth their complexity and computational cost. Finally, we compare these approaches with the systems that participated in the 300 VW competition of Shen et al. (2015).

The rest of the paper is organised as follows. Sect. 2 presents a survey of the current literature on both rigid and deformable face tracking. In Sect. 3, we present the current state-of-the-art methodologies for deformable face tracking. Since, modern face tracking consists of various modules, including face detection, model free tracking and facial landmark localisation, Sects. 3.1–3.3 briefly outline the state-of-the-art in each of these domains. Experimental results are presented in Sect. 4. Finally, in Sect. 5 we discuss the challenges that still remain to be addressed, provide future research directions and draw conclusions.

2 Related Work

Rigid and deformable tracking of faces and facial features have been a very popular topic of research over the past twenty years (Black and Yacoob 1995; Lanitis et al. 1995; Sobottka and Pitas 1996; Essa et al. 1996, 1997; Oliver et al. 1997; Decarlo and Metaxas 2000; Jepson et al. 2003; Matthews and Baker 2004; Matthews et al. 2004; Xiao et al. 2004; Patras and Pantic 2004; Kim et al. 2008; Ross et al. 2008; Papandreou and Maragos 2008; Amberg et al. 2009; Kalal et al. 2010a; Koelstra et al. 2010; Tresadern et al. 2012; Tzimiropoulos and Pantic 2013; Xiong and De la Torre 2013; Liwicki et al. 2013; Smeulders et al. 2014; Asthana et al. 2014; Tzimiropoulos and Pantic 2014; Li et al. 2016a; Xiong and De la Torre 2015; Snape et al. 2015; Wu et al. 2015; Tzimiropoulos 2015). In this section we provide an overview of face tracking spanning over the past twenty years up to the present day. In particular, we will outline the methodologies regarding rigid 2D/3D face tracking, as well as deformable 2D/3D face tracking using a monocular camera.[2] Finally, we

[1] The results and dataset of the 300 VW Challenge by Shen et al. (2015) can be found at http://ibug.doc.ic.ac.uk/resources/300-VW/ This is the first facial landmark tracking challenge on challenging long-term sequences.

[2] The problem of face tracking using commodity depth cameras, which has received a lot of attention (Göktürk and Tomasi 2004; Cai et al. 2010; Weise et al. 2011), falls outside the scope of this paper.

outline the benchmarks for both rigid and deformable face tracking.

2.1 Prior Art

The first methods for rigid 2D tracking generally revolved around the use of various features or transformations and mainly explored various color-spaces for robust tracking (Crowley and Berard 1997; Bradski 1998b; Qian et al. 1998; Toyama 1998; Jurie 1999; Schwerdt and Crowley 2000; Stern and Efros 2002; Vadakkepat et al. 2008). The general methods of choice for tracking were Mean Shift and variations such as the Continuously Adaptive Mean Shift (Camshift) algorithm (Bradski 1998a; Allen et al. 2004). The Mean Shift algorithm is a non-parametric technique that climbs the gradient of a probability distribution to find the nearest dominant mode (peak) (Comaniciu and Meer 1999; Comaniciu et al. 2000). Camshift is an adaptation of the Mean Shift algorithm for object tracking. The primary difference between CamShift and Mean Shift is that the former uses continuously adaptive probability distributions (i.e., distributions that may be recomputed for each frame) while the latter is based on static distributions, which are not updated unless the target experiences significant changes in shape, size or color. Other popular methods of choice for tracking are linear and non-linear filtering techniques including Kalman filters, as well as methodologies that fall in the general category of particle filters (Del Moral 1996; Gordon et al. 1993), such as the popular Condensation algorithm by Isard and Blake (1998). Condensation is the application of Sampling Importance Resampling (SIR) estimation by Gordon et al. (1993) to contour tracking. A recent successful 2D rigid tracker that updates the appearance model of the tracked face was proposed in Ross et al. (2008). The algorithm uses incremental Principal Component Analysis (PCA) (Levey and Lindenbaum 2000) to learn a statistical model of the appearance in an on-line manner and contrary to other eigentrackers, such as Black and Jepson (1998), it does not contain any training phase. The method in Ross et al. (2008) uses a variant of the Condensation algorithm to model the distribution over the objects location as it evolves over time. The method has initiated a line of research on robust incremental object tracking including the works of Liwicki et al. (2012b, 2013, 2012a, 2015). Rigid 3D tracking has also been studied by using generic 3D models of the face (Malciu and Prěteux 2000; La Cascia et al. 2000). For example, La Cascia et al. (2000) formulate the tracking task as an image registration problem in the cylindrically unwrapped texture space and Sung et al. (2008) combine active appearance models (AAMs) with a cylindrical head model for robust recovery of the global rigid motion. Currently, rigid face tracking is generally treated along the same lines as general model free object tracking (Jepson et al. 2003; Smeulders et al. 2014; Liwicki et al. 2013, 2012b; Ross et al. 2008; Wu et al. 2015; Li et al. 2016a). An overview of model free object tracking is given in Sect. 3.2.

Non-rigid (deformable) tracking of faces is important in many applications, spanning from facial expression analysis to motion capture for graphics and game design. Deformable tracking of faces can be further subdivided into i) tracking of certain facial landmarks (Lanitis et al. 1995; Black and Yacoob 1995; Sobottka and Pitas 1996; Xiao et al. 2004; Matthews and Baker 2004; Matthews et al. 2004; Patras and Pantic 2004; Papandreou and Maragos 2008; Amberg et al. 2009; Tresadern et al. 2012; Xiong and De la Torre 2013; Asthana et al. 2014; Xiong and De la Torre 2015) or ii) tracking/estimation of dense facial motion (Essa et al. 1996; Yacoob and Davis 1996; Essa et al. 1997; Decarlo and Metaxas 2000; Koelstra et al. 2010; Snape et al. 2015). The latter category of estimating a dense facial motion through a model-based system was proposed by MIT Media lab in mid 1990’s (Essa et al. 1997, 1996, 1994; Basu et al. 1996). In particular, the method by Essa and Pentland (1994) tracks facial motion using optical flow computation coupled with a geometric and a physical (muscle) model describing the facial structure. This modeling results in a time-varying spatial patterning of facial shape and a parametric representation of the independent muscle action groups which is responsible for the observed facial motions. In Essa et al. (1994) the physically-based face model of Essa and Pentland (1994) is driven by a set of responses from a set of templates that characterise facial regions. Model generated flow has been used by the same group in Basu et al. (1996) for motion regularisation. 3D motion estimation using sparse 3D models and optical flow estimation has also been proposed by Li et al. (1993), Bozdaği et al. (1994). Dense facial motion tracking is performed in Decarlo and Metaxas (2000) by solving a model-based (using a facial deformable model) least-squares optical flow problem. The constraints are relaxed by the use of a Kalman filter, which permits controlled constraint violations based on the noise present in the optical flow information, and enables optical flow and edge information to be combined more robustly and efficiently. Free-form deformations (Rueckert et al. 1999) are used in Koelstra et al. (2010) for extraction of dense facial motion for facial action unit recognition. Recently, Snape et al. (2015) proposed a statistical model of the facial flow for fast and robust dense facial motion extraction.

Arguably, the category of deformable tracking that has received the majority of attention is that of tracking a set of sparse facial landmarks. The landmarks are either associated to a particular sparse facial model, i.e. the popular Candide facial model by Li et al. (1993), or correspond to fiducial facial regions/parts (e.g., mouth, eyes, nose etc.) (Cootes et al. 2001). Even earlier attempts such as Essa and Pentland (1994) understood the usefulness of tracking facial regions/landmarks in order to perform robust fitting of complex facial models (currently the vast majority of dense 3D facial model tracking techniques, such as Wei et al. (2004), Zhang et al. (2008), Amberg (2011), rely on the robust tracking of a set of facial landmarks). Early approaches for tracking facial landmarks/regions included: (i) the use of templates built around certain facial regions (Essa and Pentland 1994), (ii) the use of facial classifiers to detect landmarks (Colmenarez et al. 1999) where tracking is performed using modal analysis (Tao and Huang 1998) or (iii) the use of face and facial region segmentation to detect the features where tracking is performed using block matching (Sobottka and Pitas 1996). Currently, deformable face tracking has converged with the problem of facial landmark localisation on static images. That is, the methods generally rely on fitting generative or discriminative statistical models of appearance and 2D/3D sparse facial shape at each frame. Arguably, the most popular methods are generative and discriminative variations of Active Appearance Models (AAMs) and Active Shape Models (ASMs) (Pighin et al. 1999; Cootes et al. 2001; Dornaika and Ahlberg 2004; Xiao et al. 2004; Matthews and Baker 2004; Dedeoğlu et al. 2007; Papandreou and Maragos 2008; Amberg et al. 2009; Saragih et al. 2011; Xiong and De la Torre 2013, 2015). The statistical models of appearance and shape can either be generic as in Cootes et al. (2001), Matthews and Baker (2004), Xiong and De la Torre (2013) or incrementally updated in order to better capture the face at hand, as in Sung and Kim (2009), Asthana et al. (2014). The vast majority of the facial landmark localisation methodologies require an initialisation provided by a face detector. More details regarding current state-of-the-art in facial landmark localisation can be found in Sect. 3.3.

Arguably, the current practise regarding deformable face tracking includes the combination of a generic face detection and generic facial landmark localisation technique (Saragih et al. 2011; Xiong and De la Torre 2013, 2015; Alabort-i-Medina and Zafeiriou 2015; Asthana et al. 2015). For example, popular approaches include successive application of the face detection and facial landmark localisation procedure at each frame. Another approach performs face detection in the first frame and then applies facial landmark localisation at each consecutive frame using the fitting result of the previous frame as initialisation. Face detection can be re-applied in case of failure. This is the approach that is used by popular packages such as Asthana et al. (2014). In this paper, we thoroughly evaluate variations of the above approaches. Furthermore, we consider the use of modern model free state-of-the-art trackers for rigid 2D tracking in order to be used as initialisation for the facial landmark localisation procedure. This is pictorially described in Fig. 1.

2.2 Face Tracking Benchmarking

For assessing the performance of rigid 2D face tracking several short face sequences have been annotated with regards to the facial region (using a bounding box style annotation). One of the first sequences that has been annotated for this task is the so-called Dudek sequence by Ross et al. (2015).[3] Nowadays, several such sequences have been annotated and are publicly available, such as the ones by Liwicki et al. (2016), Li et al. (2016b), Wu et al. (2015).

The performance of deformable dense facial tracking methodologies was usually assessed by using markers (Decarlo and Metaxas 2000), simulated data (Snape et al. 2015), visual inspection (Decarlo and Metaxas 2000; Essa et al. 1997, 1996; Yacoob and Davis 1996; Snape et al. 2015; Koelstra et al. 2010) or indirectly by the use of the dense facial motion for certain tasks, such as expression analysis (Essa et al. 1996; Yacoob and Davis 1996; Koelstra et al. 2010). Regarding tracking of facial landmarks, up until recently, the preferred method for assessing the performance was visual inspection in a number of selected facial videos (Xiong and De la Torre 2013; Tresadern et al. 2012). Other methods were assessed on a small number of short (a few seconds in length) annotated facial videos (Sagonas et al. 2014; Asthana et al. 2014). Until recently the longest annotated facial video sequence was the so-called talking face of Cootes (2016) which was used to evaluate many tracking methods including Orozco et al. (2013), Amberg et al. (2009). The talking face video comprises of 5000 frames (around 200 seconds) taken from a video of a person engaged in a conversation. The talking face video was initially tracked using an Active Appearance Model (AAM) that had a shape model and a total of 68 landmarks are provided. The tracked landmarks were visually checked and manually corrected where necessary.

Recently, Xiong and De la Torre (2015) introduced a benchmark for facial landmark tracking using videos from the Distracted Driver Face (DDF) and Naturalistic Driving Study (NDS) in Campbell (2016).[4] The DDF dataset contains 15 sequences with a total of 10,882 frames. Each sequence displays a single subject posing as the distracted driver in a stationary vehicle or indoor environment. 12 out of 15 videos were recorded with subjects sitting inside of a vehicle. Five of them were recorded during the night under infrared (IR) light and the rest were recorded during the daytime under natural lighting. The remaining three were recorded indoors. The NDS database contains 20 sub-sequences of

[3] The Dudek sequence has been annotated with regards to certain facial landmarks only to be used for the estimation of an affine transformation.

[4] In a private communication, the authors of Xiong and De la Torre (2015) informed us that the annotated data, as described in the paper, will not be made publicly available (at least not in the near future).

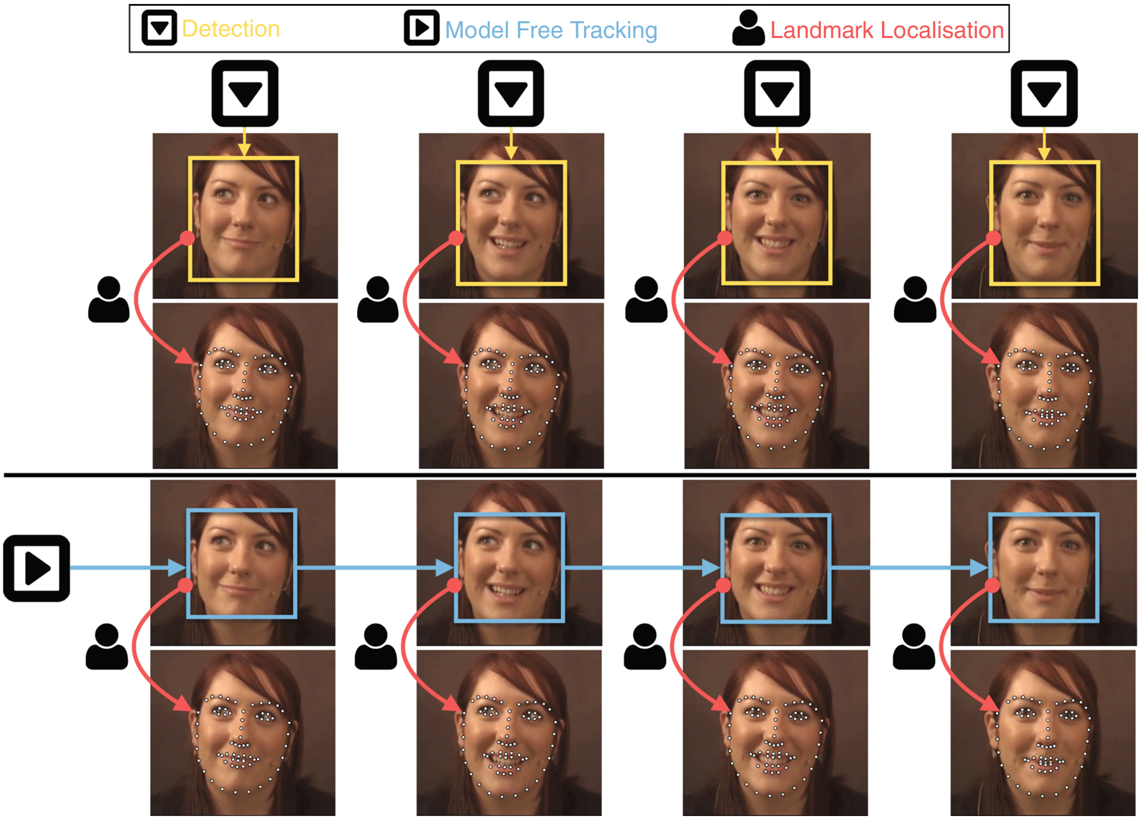


Fig. 1 Overview of the standard approaches for deformable face tracking. *(Top)* face detection is applied independently at each frame of the video followed by facial landmark localisation. *(Bottom)* model free tracking is employed, initialised with the bounding box of the face at the first frame, followed by facial landmark localisation

driver faces recorded during a drive conducted between the Blacksburg, VA and Washington, DC areas (NDS is more challenging than DDF since its videos are of lower spatial and temporal resolution). Each video of the NDS database has one minute duration recorded at 15 frames per second (fps) with a 360 × 240 resolution. For both datasets one in every ten frames was annotated using either 49 landmarks for near-frontal faces or 31 landmarks for profile faces. The database contains many extreme facial poses (90° yaw, 50∘ pitch) as well as many faces under extreme lighting condition (e.g., IR). In total the dataset presented in Xiong and De la Torre (2015) contains between 2000 to 3000 annotated faces (please refer to Xiong and De la Torre (2015) for exemplar annotations).

The only existing large in-the-wild benchmark for facial landmark tracking was recently introduced by Shen et al. (2015). The benchmark consists of 114 videos with varying difficulty and provides annotations generated in a semi-automatic manner (Chrysos et al. 2015; Shen et al. 2015; Tzimiropoulos 2015). This challenge, called 300 VW, is the only existing large-scale comprehensive benchmark for deformable model tracking. More details regarding the dataset of the 300 VW benchmark can be found in Sect. 4.1. The performance of the pipelines considered in this paper are compared with the participating methods of the 300 VW challenge in Sect. 4.8.

3 Deformable Face Tracking

In this paper, we focus on the problem of performing deformable face tracking across long-term sequences within unconstrained videos. The problem of tracking across long-term sequences is particularly challenging as the appearance of the face may change significantly during the sequence due to occlusions, illumination variation, motion artifacts and head pose. For the problem of deformable tracking, however, the problem is further complicated by the expectation of recovering a set of accurate fiducial points in conjunction with successfully tracking the object. As described in Sect. 2, current deformable facial tracking methods mainly concentrate on performing face detection per frame and then performing facial landmark localisation. However, we consider the most important metric for measuring the success of deformable face tracking as the facial landmark localisation accuracy. Given this, there are a number of strategies that could feasibly be employed in order to attempt to minimise the total facial landmark localisation error across the entire sequence. Therefore, we take advantage of current advances in face detection, model free tracking and facial landmark localisation techniques in order to perform deformable face tracking. Specifically, we investigate three strategies for deformable tracking:

1. **Detection + landmark localisation** Face Detection per frame, followed by facial landmark localisation initialised within the facial bounding boxes. This scenario is visualised in Fig. 1 (top).
2. **Model free tracking + landmark localisation** Model free tracking, initialised around the interior of the face within the first frame, followed by facial landmark localisation within the tracked box. This scenario is visualised in Fig. 1 (bottom).
3. **Hybrid systems** Hybrid methods that attempt to improve the robustness of the placement of the bounding box for landmark localisation. Namely, we investigate methods for failure detection, trajectory smoothness and reinitialisation. Examples of such methods are pictorially demonstrated in Figs. 4 and 8.

Note that we focus on combinations of methods that provide bounding boxes of the facial region followed by landmark localisation. This is due to the fact that the current set of state-of-the-art landmark localisation methods are all local methods and require initialisation within the facial region. Although joint face detection and landmark localisation methods have been proposed (Zhu and Ramanan 2012; Chen et al. 2014), they are not competitive with the most recent set of landmark localisation methods. For this reason, in this paper we focus on the combination of bounding box estimators with state-of-the-art local landmark localisation techniques.

The remainder of this Section will give a brief overview of the literature concerning face detection, model free tracking and facial landmark localisation.

3.1 Face Detection

Face detection is among the most important and popular tasks in Computer Vision and an essential step for applications such as face recognition and face analysis. Although it is one of the oldest tasks undertaken by researchers (the early works appeared about 45 years ago (Sakai et al. 1972; Fischler and Elschlager 1973)), it is still an open and challenging problem. Recent advances can achieve reliable performance under moderate illumination and pose conditions, which led to the installation of simple face detection technologies in everyday devices such as digital cameras and mobile phones. However, recent benchmarks (Jain and Learned-Miller 2010) show that the detection of faces on arbitrary images is still a very challenging problem.

Since face detection has been a research topic for so many decades, the existing literature is, naturally, extremely extensive. The fact that all recent face detection surveys (Hjelmås and Low 2001; Yang et al. 2002; Zhang and Zhang 2010; Zafeiriou et al. 2015) provide different categorisations of the relative literature is indicative of the huge range of existing techniques. Consequently, herein, we only present a basic outline of the face detection literature. For an extended review, the interested reader may refer to the most recent face detection survey in Zafeiriou et al. (2015).

According to the most recent literature review Zafeiriou et al. (2015), existing methods can be separated in two major categories. The first one includes methodologies that learn a set of rigid templates, which can be further split in the following groups: (i) boosting-based methods, (ii) approaches that utilise SVM classifiers, (ii) exemplar-based techniques, and (iv) frameworks based on Neural Networks. The second major category includes deformable part models, i.e. methodologies that learn a set of templates per part as well as the deformations between them.

Boosting Methods Boosting combines multiple "weak" hypotheses of moderate accuracy in order to determine a highly accurate hypothesis. The most characteristic example is Adaptive Boosting (AdaBoost) which is utilised by the most popular face detection methodology, i.e. the Viola–Jones (VJ) detector of Viola and Jones (2001, 2004). Characteristic examples of other methods that employ variations of AdaBoost include Li et al. (2002), Wu et al. (2004), Mita et al. (2005). The original VJ algorithm used Haar features, however boosting (or cascade of classifiers methodologies in general) have been shown to greatly benefit from robust features (Köstinger et al. 2012; Jun et al. 2013; Li et al. 2011; Li and Zhang 2013; Mathias et al. 2014; Yang et al. 2014a), such as HOG (Dalal and Triggs 2005), SIFT (Lowe 1999), SURF (Bay et al. 2008) and LBP (Ojala et al. 2002). For example, SURF features have been successfully combined with a cascade of weak classifiers in Li et al. (2011), Li and Zhang (2013), achieving faster convergence. Additionally, Jun et al. (2013) propose robust face specific features that combine both LBP and HOG. Mathias et al. (2014) recently proposed an approach (so called HeadHunter) with state-of-the-art performance that employs various robust features with boosting. Specifically, they propose the adaptation of Integral Channel Features (ICF) (Dollár et al. 2009) with HOG and LUV colour channels, combined with global feature normalisation. A similar approach is followed by Yang et al. (2014a), in which they combine gray-scale, RGB, HSV, LUV, gradient magnitude and histograms within a cascade of weak classifiers.

SVM Classifiers Maximum margin classifiers, such as Support Vector Machines (SVMs), have become popular for face detection (Romdhani et al. 2001; Heisele et al. 2003; Rätsch et al. 2004; King 2015). Even though their detection speed was initially slow, various schemes have been proposed to speed up the process. Romdhani et al. (2001) propose a method that computes a reduced set of vectors from the original support vectors that are used sequentially in order to make early rejections. A similar approach is adopted by Rätsch et al. (2004). A hierarchy of SVM classifiers trained on different resolutions is applied in Heisele et al. (2003). King (2015) proposes an algorithm for efficient learning of a max-margin classifier using all the sub-windows of the training images, without applying any sub-sampling, and formulates a convex optimisation that finds the global optimum. Moreover, SVM classifiers have also been used for multi-view face detection (Li et al. 2000; Wang and Ji 2004). For example, Li et al. (2000) first apply a face pose estimator based on support vector regression (SVR), followed by an SVM face detector for each pose.

Exemplar-Based Techniques These methods aim to match a test image against a large set of facial images. This approach is inspired by principles used in image retrieval and requires that the exemplar set covers the large appearance variation of human face. Shen et al. (2013) employ bag-of-word image retrieval methods to extract features from each exemplar, which creates a voting map for each exemplar that functions as a weak classifier. Thus, the final detection is performed by combining the voting maps. A similar methodology is applied in Li et al. (2014), with the difference that specific exemplars are used as weak classifiers based on a boosting strategy. Recently, Kumar et al. (2015) proposed an approach that enhances the voting procedure by using semantically related visual words as well as weighted occurrence of visual words based on their spatial distributions.

Convolutional Neural Networks Another category, similar to the previous rigid template-based ones, includes the employment of Convolutional Neural Networks (CNNs) and Deep CNNs (DCNNs) (Osadchy et al. 2007; Zhang and Zhang 2014; Ranjan et al. 2015; Li et al. 2015a; Yang et al. 2015b). Osadchy et al. (2007) use a network with four convolution layers and one fully connected layer that rejects the non-face hypotheses and estimates the pose of the correct face hypothesis. Zhang and Zhang (2014) propose a multi-view face detection framework by employing a multi-task DCNN for face pose estimation and landmark localization in order to obtain better features for face detection. Ranjan et al. (2015) combine deep pyramidal features with Deformable Part Models. Recently, Yang et al. (2015b) proposed a DCNN architecture that is able to discover facial parts responses from arbitrary uncropped facial images without any part supervision and report state-of-the-art performance on current face detection benchmarks.

Deformable Part Models DPMs (Schneiderman and Kanade 2004; Felzenszwalb and Huttenlocher 2005; Felzenszwalb et al. 2010; Zhu and Ramanan 2012; Yan et al. 2013; Li et al. 2013a; Yan et al. 2014; Mathias et al. 2014; Ghiasi and Fowlkes 2014; Barbu et al. 2014) learn a patch expert for each part of an object and model the deformations between parts using spring-like connections based on a tree structure. Consequently, they perform joint facial landmark localisation and face detection. Even though they are not the best performing methods for landmark localisation, they are highly accurate for face detection in-the-wild. However, their main disadvantage is their high computational cost. Pictorial Structures (PS) (Fischler and Elschlager 1973; Felzenszwalb and Huttenlocher 2005) are the first family of DPMs that appeared. They are generative DPMs that assume Gaussian distributions to model the appearance of each part, as well as the deformations. They became a very popular line of research after the influential work in Felzenszwalb and Huttenlocher (2005) that proposed a very efficient dynamic programming algorithm for finding the global optimum based on Generalized Distance Transform. Many discriminatively trained DPMs (Felzenszwalb et al. 2010; Zhu and Ramanan 2012; Yan et al. 2013, 2014) appeared afterwards, which learn the patch experts and deformation parameters using discriminative classifiers, such as latent SVM.

DPMs can be further separated with respect to their training scenario into: (i) weakly supervised and (ii) strongly supervised. Weakly-supervised DPMs (Felzenszwalb et al. 2010; Yan et al. 2014) are trained using only the bounding boxes of the positive examples and a set of negative examples. The most representative example is the work by Felzenszwalb et al. (2010), which has proved to be very efficient for generic object detection. Under a strongly supervised scenario, it is assumed that a training database with images annotated with figucial landmarks is available. Several strongly supervised methods exist in the literature (Felzenszwalb and Huttenlocher 2005; Zhu and Ramanan 2012; Yan et al. 2013; Ghiasi and Fowlkes 2014). Ghiasi and Fowlkes (2014) propose an hierarchical DPM that explicitly models parts' occlusions. In Zhu and Ramanan (2012) it is shown that a strongly supervised DPM outperforms, by a large margin, a weakly supervised one. In contrast, HeadHunter by Mathias et al. (2014) shows that a weakly supervised DPM can outperform all current state-of-the-art face detection methodologies including the strongly supervised DPM of Zhu and Ramanan (2012).

According to FDDB (Jain and Learned-Miller 2010), which is the most well established face detection benchmark, the currently top-performing methodology is the one by Ranjan et al. (2015), which combines DCNNs with a DPM. Some of the top-performing systems consist of commercial software, thus we did use the deep methods of Hu and Ramanan (2016), Zhang et al. (2016) that are available as open source

Table 1 The set of detectors used in this paper

Method	Citation(s)	Rigid template	DPM	Implementation
DPM	Felzenszwalb et al. (2010)		✓	https://github.com/menpo/ffld2
	Mathias et al. (2014)			
	Alabort-i-Medina et al. (2014)			
HR-TF	Hu and Ramanan (2016)	✓		https://www.cs.cmu.edu/~peiyunh/tiny/
MTCNN	Zhang et al. (2016)	✓		https://goo.gl/4BMGeR
NPD	Liao et al. (2016)	✓		https://goo.gl/dRXp8d
SS-DPM	Zhu and Ramanan (2012)		✓	https://www.ics.uci.edu/~xzhu/face
SVM+HOG	King (2015)	✓		https://github.com/davisking/dlib
	King (2009)			
VJ	Viola and Jones (2004)	✓		http://opencv.org
	Bradski (2000)			
VPHR	Kumar et al. (2015)	✓		http://cvit.iiit.ac.in/projects/exemplar/

The table reports the short name of the method, the relevant citation(s) as well as the link to the implementation used

with the method of Hu and Ramanan (2016) reporting the latest best performance in FDDB. Additionally, we employ the top performing SVM-based method for learning rigid templates (King 2015), the best weakly and strongly supervised DPM implementations of Mathias et al. (2014) and Zhu and Ramanan (2012), along with the best performing exemplar-based technique of Kumar et al. (2015) . Finally, we also use the popular VJ algorithm (Viola and Jones 2001, 2004) as a baseline face detection method. The employed face detection implementations are summarised in Table 1.

3.2 Model Free Tracking

Model free tracking is an extremely active area of research. Given the initial state (e.g., position and size of the containing box) of a target object in the first image, model free tracking attempts to estimate the states of the target in subsequent frames. Therefore, model free tracking provides an excellent method of initialising landmark localisation methods.

The literature on model free tracking is vast. For the rest of this section, we will provide an extremely brief overview of model free tracking that focuses primarily on areas that are relevant to the tracking methods we investigated in this paper. We refer the interested reader to the wealth of tracking surveys (Li et al. 2013b; Smeulders et al. 2014; Salti et al. 2012; Yang et al. 2011) and benchmarks (Wu et al. 2013, 2015; Kristan et al. 2013, 2014, 2015, 2016; Smeulders et al. 2014) for more information on model free tracking methods.

Generative Trackers These trackers attempt to model the objects appearance directly. This includes template based methods, such as those by Matthews et al. (2004), Baker and Matthews (2004), Sevilla-Lara and Learned-Miller (2012), as well as parametric generative models such as Balan and Black (2006), Ross et al. (2008), Black and Jepson (1998) , Xiao et al. (2014). The work of Ross et al. (2008) introduces online subspace learning for tracking with a sample mean update, which allows the tracker to account for changes in illumination, viewing angle and pose of the object. The idea is to incrementally learn a low-dimensional subspace and adapt the appearance model on object changes. The update is based on an incremental principal component analysis (PCA) algorithm, however it seems to be ineffective at handling large occlusions or non-rigid movements due to its holistic model. To alleviate the partial occlusion, Xiao et al. (2014) suggest the use of square templates along with PCA. Another popular area of generative tracking is the use of sparse representations for appearance. In Mei and Ling (2011), a target candidate is represented by a sparse linear combination of target and trivial templates. The coefficients are extracted by solving an ℓ_1 minimisation problem with non-negativity constraints, while the target templates are updated online. However, solving the ℓ_1 minimisation for each particle is computationally expensive. A generalisation of this tracker is the work of Zhang et al. (2012), which learns the representation for all particles jointly. It additionally improves the robustness by exploiting the correlation among particles. An even further abstraction is achieved in Zhang et al. (2014d) where a low-rank sparse representation of the particles is encouraged. In Zhang et al. (2014c), the authors generalise the low-rank constraint of Zhang et al. (2014d) and add a sparse error term in order to handle outliers. Another low-rank formulation was used by Wu et al. (2012) which is an online version of the RASL (Peng et al. 2012) algorithm and attempts to jointly align the input sequence using convex optimisation.

Keypoint Trackers These trackers (Pernici and Del Bimbo 2014; Poling et al. 2014; Hare et al. 2012; Nebehay and

Pflugfelder 2015) attempt to use the robustness of keypoint detection methodologies like SIFT (Lowe 1999) or SURF (Bay et al. 2008) in order to perform tracking. Pernici and Del Bimbo (2014) collected multiple descriptors of weakly aligned keypoints over time and combined these matched keypoints in a RANSAC voting scheme. Nebehay and Pflugfelder (2015) utilises keypoints to vote for the object center in each frame. A consensus-based scheme is applied for outlier detection and the votes are transformed based on the current key point arrangement to consider scale and rotation. However, keypoint methods may suffer from difficulty in capturing the global information of the tracked target by only considering the local points.

Discriminative Trackers These trackers attempt to explicitly model the difference between the object appearance and the background. Most commonly, these methods are named "tracking-by-detection" techniques as they involve classifying image regions as either part of the object or the background. In their work, Grabner et al. (2006) propose an online boosting method to select and update discriminative features which allows the system to account for minor changes in the object appearance. However, the tracker fails to model severe changes in appearance. Babenko et al. (2011) advocate the use of a multiple instance learning boosting algorithm to mitigate the drifting problem. More recently, discriminative correlation filters (DCF) have become highly successful at tracking. The DCF is trained by performing a circular sliding window operation on the training samples. This periodic assumption enables efficient training and detection by utilizing the Fast Fourier Transform (FFT). Danelljan et al. (2014) learn separate correlation filters for the translation and the scale estimation. In Danelljan et al. (2015), the authors introduce a sparse spatial regularisation term to mitigate the artifacts at the boundaries of the circular correlation. In contrast to the linear regression commonly used to learn DCFs, Henriques et al. (2015) apply a kernel regression and propose its multi-channel extension to enable to the use of features such as HOG Dalal and Triggs (2005). Li et al. (2015b) propose a new use for particle filters in order to choose reliables patches to consider part of the object. These patches are modelled using a variant of the method proposed by Henriques et al. (2015). Hare et al. (2011) propose the use of structured output prediction. By explicitly allowing the outputs to parametrize the needs of the tracker, an intermediate classification step is avoided.

Part-based Trackers These trackers attempt to implicitly model the parts of an object in order to improve tracking performance. Adam et al. (2006) represent the object with multiple arbitrary patches. Each patch votes on potential positions and scales of the object and a robust statistic is employed to minimise the voting error. Kalal et al. (2010b) sample the object and the points are tracked independently in each frame by estimating optical flow. Using a forward–backward measure, the erroneous points are identified and the remaining reliable points are utilised to compute the optimal object trajectory. Yao et al. (2013) adapt the latent SVM of Felzenszwalb et al. (2010) for online tracking, by restricting the search in the vicinity of the location of the target object in the previous frame. In comparison to the weakly supervised part-based model of Yao et al. (2013), in Zhang and van der Maaten (2013) the authors recommend an online strongly supervised part-based deformable model that learns the representation of the object and the representation of the background by training a classifier. Wang et al. (2015) employ a part-based tracker by estimating a direct displacement prediction of the object. A cascade of regressors is utilised to localise the parts, while the model is updated online and the regressors are initialised by multiple motion models at each frame.

Given the wealth of available trackers, selecting appropriate trackers for deformable tracking purposes poses a difficult proposition. In order to attempt to give as broad an overview as possible, we selected trackers from each of the aforementioned categories. Therefore, in this paper we compare against 27 trackers which are outlined in Table 2. SRDCF (Danelljan et al. 2015), KCF (Henriques et al. 2015), LCT (Ma et al. 2015), STAPLE (Bertinetto et al. 2016a) and DSST (Danelljan et al. 2014) are all discriminative trackers based on DCFs. They all performed well in the VOT 2015 (Kristan et al. 2015) challenge and DSST was the winner of VOT 2014 (Kristan et al. 2014). The trackers of Danelljan et al. (2016), Qi et al. (2016); Nam and Han (2016), Bertinetto et al. (2016b) are indicative trackers that employ neural networks and achieve top results. STRUCK (Hare et al. 2011) is a discriminative tracker that performed very well in the Online Object Tracking benchmark (Wu et al. 2013), while the more recent method of Ning et al. (2016) improves the computational burden of the structural SVM of STRUCK and reports superior results. SPOT (Zhang and van der Maaten 2014) is a strong performing part based tracker, CMT (Nebehay and Pflugfelder 2015) is a strong performing keypoint based tracker, LRST (Zhang et al. 2014d) and ORIA (Wu et al. 2012) are recent generative trackers. RPT (Li et al. 2015b) is a recently proposed technique that reported state-of-the-art results on the Online Object Tracking benchmark (Wu et al. 2013). TLD (Kalal et al. 2012), MIL (Babenko et al. 2011), FCT (Zhang et al. 2014c), DF (Sevilla-Lara and Learned-Miller 2012) and IVT (Ross et al. 2008) were included as baseline tracking methods with publicly available implementations. Finally, the CAMSHIFT and PF methods (Bradski 1998a; Isard and Blake 1996) are included as very influential trackers used in the previous decades for tracking.

Table 2 The set of trackers that are used in this paper

Method	Citation(s)	D	G	P	K	NN	Implementation
CAMSHIFT	Bradski (1998a)	✓					http://opencv.org
CCOT	Danelljan et al. (2016)	✓				✓	https://goo.gl/Rnf73K
CMT	Nebehay and Pflugfelder (2015)				✓		https://github.com/gnebehay/CppMT
DF	Sevilla-Lara and Learned-Miller (2012)		✓				http://goo.gl/YmG6W4
DLSSVM	Ning et al. (2016)	✓					https://goo.gl/m4ro8x
DSST	Danelljan et al. (2014) King (2009)	✓					https://github.com/davisking/dlib
FCT	Zhang et al. (2014c)	✓	✓				http://goo.gl/Ujc5B0
HDT	Qi et al. (2016)					✓	https://goo.gl/9KgteR
IVT	Ross et al. (2008)		✓				http://goo.gl/WtbOIX
KCF	Henriques et al. (2015)	✓					https://github.com/joaofaro/KCFcpp
LCT	Ma et al. (2015)	✓					https://goo.gl/8kaO7T
LRST	Zhang et al. (2014d)		✓				http://goo.gl/ZC9JbQ
MDNET	Nam and Han (2016)	✓				✓	https://github.com/HyeonseobNam/MDNet
MEEM	Zhang et al. (2014a)	✓					https://goo.gl/Bj6typ
MIL	Babenko et al. (2011) Bradski (2000)	✓					http://opencv.org
ORIA	Wu et al. (2012)		✓				https://goo.gl/RT3zNC
PF	Isard and Blake (1996)		✓				https://goo.gl/tSZcAg
RPT	Li et al. (2015b)	✓					https://github.com/ihpdep/rpt
SIAM-OXF	Bertinetto et al. (2016b)	✓				✓	https://goo.gl/sjGgVj
SPOT	Zhang and van der Maaten (2014)	✓		✓			http://visionlab.tudelft.nl/spot
SPT	Yang et al. (2014b)	✓					https://goo.gl/EOquai
SRDCF	Danelljan et al. (2015)	✓					https://goo.gl/Q9d1O5
STAPLE	Bertinetto et al. (2016a)	✓					https://github.com/bertinetto/staple
STCL	Zhang et al. (2014b)	✓					https://goo.gl/l29dQg
STRUCK	Hare et al. (2011)	✓					http://goo.gl/gLR93b
TGPR	Gao et al. (2014)	✓					https://goo.gl/EBw0WI
TLD	Kalal et al. (2012)	✓					https://github.com/zk00006/OpenTLD

The table reports the short name of the method, the relevant citation(s) as well as the link to the implementation used. The initials stand for: (*D*)iscriminative, (*G*)enerative, (*P*)art-based, (*K*)eypoint trackers, and *NN* for trackers that employ neural networks

3.3 Facial Landmark Localisation

Statistical deformable models have emerged as an important research field over the last few decades, existing at the intersection of computer vision, statistical pattern recognition and machine learning. Statistical deformable models aim to solve generic object alignment in terms of localisation of fiducial points. Although deformable models can be built for a variety of object classes, the majority of ongoing research has focused on the task of facial alignment. Recent large-scale challenges on facial alignment (Sagonas et al. 2013b, a, 2015) are characteristic examples of the rapid progress being made in the field.

Currently, the most commonly-used and well-studied face alignment methods can be separated into two major families: (i) *discriminative* models that employ regression in a cascaded manner, and (ii) *generative* models that are iteratively optimised.

Regression-Based Models The methodologies of this category aim to learn a regression function that regresses from the object's appearance (e.g. commonly handcrafted features) to the target output variables (either the landmark coordinates or the parameters of a statistical shape model). Although the history behind using linear regression in order to tackle the problem of face alignment spans back many years (Cootes et al. 2001), the research community turned towards alternative approaches due to the lack of sufficient data for training accurate regression functions. Nevertheless, recently regression-based techniques have prevailed in the field thanks to the wealth of annotated data and effective

Table 3 The landmark localisation methods employed in this paper

Method	Citation(s)	Discriminative	Generative	Implementation
AAM	Tzimiropoulos (2015)		✓	https://github.com/menpo/menpofit
	Alabort-i-Medina et al. (2014)			
ERT	Kazemi and Sullivan (2014)	✓		https://github.com/davisking/dlib
	King (2009)			
CFSS	Zhu et al. (2015)	✓		https://github.com/zhusz/CVPR15-CFSS
SDM	Xiong and De la Torre (2013)	✓		https://github.com/menpo/menpofit
	Alabort-i-Medina et al. (2014)			

The table reports the short name of the method, the relevant citation(s) as well as the link to the implementation used

handcrafted features (Lowe 1999; Dalal and Triggs 2005). Recent works have shown that excellent performance can be achieved by employing a cascade of regression functions (Burgos-Artizzu et al. 2013; Xiong and De la Torre 2013, 2015; Dollár et al. 2010; Cao et al. 2014; Kazemi and Sullivan 2014; Ren et al. 2014; Asthana et al. 2014; Tzimiropoulos 2015; Zhu et al. 2015). Regression based methods can be approximately seperated into two categories depending on the nature of the regression function employed. Methods that employ a linear regression such as the supervised descent method (SDM) of Xiong and De la Torre (2013) tend to employ robust hand-crafted features (Xiong and De la Torre 2013; Asthana et al. 2014; Xiong and De la Torre 2015; Tzimiropoulos 2015; Zhu et al. 2015). On the other hand, methods that employ tree-based regressors such as the explicit shape regression (ESR) method of Cao et al. (2014), tend to rely on data driven features that are optimised directly by the regressor (Burgos-Artizzu et al. 2013; Cao et al. 2014; Dollár et al. 2010; Kazemi and Sullivan 2014).

Generative Models The most dominant representative algorithm of this category is, by far, the active appearance model (AAM). AAMs consist of parametric linear models of both shape and appearance of an object, typically modelled by Principal Component Analysis (PCA). The AAM objective function involves the minimisation of the appearance reconstruction error with respect to the shape parameters. AAMs were initially proposed by Cootes et al. (1995, 2001), where the optimisation was performed by a single regression step between the current image reconstruction residual and an increment to the shape parameters. However, Matthews and Baker (2004), Baker and Matthews (2004) linearised the AAM objective function and optimised it using the Gauss-Newton algorithm. Following this, Gauss-Newton optimisation has been the modern method for optimising AAMs. Numerous extensions have been published, either related to the optimisation procedure (Papandreou and Maragos 2008; Tzimiropoulos and Pantic 2013; Alabort-i-Medina and Zafeiriou 2014, 2015; Tzimiropoulos and Pantic 2014) or the model structure (Tzimiropoulos et al. 2012; Antonakos et al. 2014; Tzimiropoulos et al. 2014; Antonakos et al. 2015b, a).

In recent challenges by Sagonas et al. (2013a, 2015), discriminative methods have been shown to represent the current state-of-the-art. However, in order to enable a fair comparison between types of methods we selected a representative set of landmark localisation methods to compare with in this paper. The set of landmark localisation methods used in the paper is given in Table 3. We chose to use ERT (Kazemi and Sullivan 2014) as it is extremely fast and the implementation provided by King (2009) is the best known implementation of a tree-based regressor. We chose CFSS (Zhu et al. 2015) as it is the current state-of-the-art on the data provided by the 300W competition of Sagonas et al. (2013a). We used the Gauss-Newton Part-based AAM of Tzimiropoulos and Pantic (2014) as the top performing generative localisation method, as provided by the Menpo Project (Alabort-i-Medina et al. 2014). Finally, we also demonstrated an SDM (Xiong and De la Torre 2013) as implemented by Alabort-i-Medina et al. (2014) as a baseline.

4 Experiments

In this section, details of the experimental evaluation are established. Firstly, the datasets employed for the evaluation, training and validation are introduced in Sect. 4.1. Next, Sect. 4.2 provides details of the training procedures and of the implementations that are relevant to all experiments. Following this, in Sects. 4.3—4.7, we describe the set of experiments that were conducted in this paper, which are summarised in Table 4. Finally, experimental Sect. 4.8 compares the best results from the previous experiments to the winners of the 300 VW competition in Shen et al. (2015).

In the following sections, due to the very large amount of methodologies taken into account, we provide a summary of all the results as tables and only the top five methods as graphs for clarity. Please refer to the supplementary material for an extensive report of the experimental results. Additionally, we

Table 4 The set of experiments conducted in this paper

Experiment	Section	Tracking	Detection	Landmark localisation	Failure checking	Re-initialisation	Kalman Smoothing
1	4.3		✓	✓			
2	4.4		✓	✓		✓	
3	4.5	✓		✓			
4	4.6	✓		✓	✓	✓	
5	4.7	✓	✓	✓			✓
6	4.8	Comparison against state-of-the-art of 300 VW competition (Shen et al. 2015).					

This table is intended as an overview of the battery of experiments that were conducted, as well as providing a reference to the relevant section

provide videos with the tracking results for the experiments of Sects. 4.3, and 4.5 for qualitative comparison.[5],[6]

4.1 Dataset

All the comparisons are conducted in the testset of the 300 VW dataset collected by Shen et al. (2015). This recently introduced dataset contains 114 videos (50 for training and 64 for testing). The videos are separated into the following 3 categories:

- *Category 1* This category is composed of videos captured in well-lit environments without any occlusions.
- *Category 2* The second category includes videos captured in unconstrained illumination conditions.
- *Category 3* The final category consists of video sequences captured in totally arbitrary conditions (including severe occlusions and extreme illuminations).

Each video includes only one person and is annotated using the 68 point mark-up employed by Gross et al. (2010) and Sagonas et al. (2015) for Multi-PIE and 300W databases, respectively. All videos include between 1500 frames and 3000 frames with a large variety of expressions, poses and capturing conditions, which makes the dataset very challenging for deformable facial tracking. A number of exemplar images, which are indicative of the challenges of each category, are provided in Fig. 2. We note that, in contrast to the results of Shen et al. (2015) in the original 300 VW competition, we used the most recently provided annotations (See footnote 1) which have been corrected and do not contain missing frames. Therefore, we also provide updated results following the participants of the 300 VW competition.

The public datasets of IBUG (Sagonas et al. 2013a), HELEN (Le et al. 2012), AFW (Zhu and Ramanan 2012) and LFPW (Belhumeur et al. 2013) are employed for training all the landmark localisation methods. This is further explained in Sect. 4.2.1 below.

4.2 Implementation Details

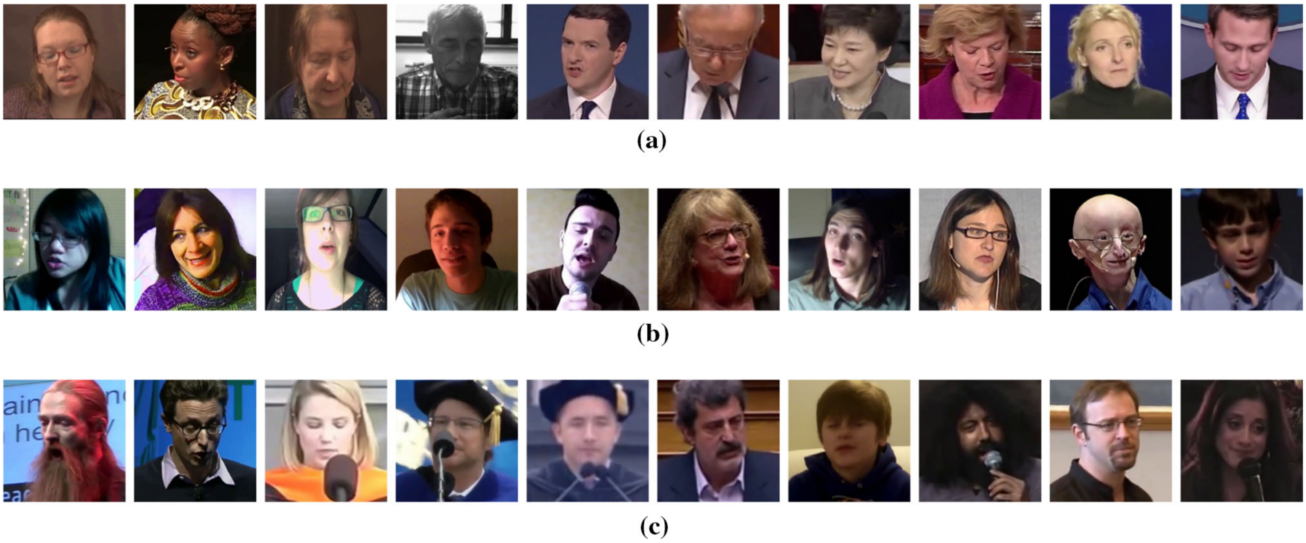
The authors' implementations are utilised for the trackers, as outlined in Table 2. Similarly, the face detectors' implementations are outlined in Table 1. HOG + SVM was provided by the Dlib project of King (2015, 2009), the Weakly Supervised DPM (DPM) (Felzenszwalb et al. 2010) was the model provided by Mathias et al. (2014) and the code of Dubout and Fleuret (2012, 2013) was used to perform the detection. Moreover, the Strongly Supervised DPM (SS-DPM) of Zhu and Ramanan (2012) was provided by the authors and, finally, the OpenCV implementation by Bradski (2000) was used for the VJ detector (Viola and Jones 2004). The default parameters were used in all cases. The pre-trained detectors' models were utilised; only the most confident detection was exported per frame, there was no effort to maximise the overlap with the ground-truth bounding box; in all videos there is only one person per frame.

For face alignment, as outlined in Table 3, the implementation of CFSS provided by Zhu et al. (2015) is adopted, while the implementations provided by Alabort-i-Medina et al. (2014) in the Menpo Project are employed for the patch-based AAM of Tzimiropoulos and Pantic (2014) and the SDM of Xiong and De la Torre (2013). Lastly, the implementation of ERT (Kazemi and Sullivan 2014) is provided by King (2009) in the Dlib library. For the three latter methods, following the original papers and the code's documentation, several parameters were validated and chosen based on the results in a validation set that consisted of a few videos from the 300 VW training set.

The details of the parameters utilised for the patch-based AAM, SDM and ERT are the following: For AAM, we used the algorithm of Tzimiropoulos and Pantic (2014) and applied a 2-level Gaussian pyramid with 4 and 10 shape com-

[5] In https://youtu.be/Lx5gHvErqX8 we provide a video with the tracking results of the top methods for face detection followed by landmark localisation (Sect. 4.3, Table 6, Fig. 3) for qualitative comparison.

[6] In https://youtu.be/SNr39MH3dh8 we provide a video with the tracking results of the top methods for model free tracking followed by landmark localisation (Sect. 4.5, Table 8, Fig. 7) for qualitative comparison.


(a)

(b)

(c)

Fig. 2 Example frames from the 300 VW dataset by Shen et al. (2015). Each *row* contains 10 exemplar images from each category, that are indicative of the challenges that characterise the videos of the category. **a** Category 1. **b** Category 2. **c** Category 3

ponents, and 60 and 150 appearance components in each scale, respectively. For the SDM, a 4-level Gaussian pyramid was employed. SIFT (Lowe 1999) feature vectors of length 128 were extracted at the first 3 scales, using RootSIFT by Arandjelović and Zisserman (2012). Raw pixel intensities were used at the highest scale.

Part of the experiments was conducted on the cloud software of Koukis et al. (2013) and the web application of Pérez and Granger (2007), while the rest of the functionality was provided by the Python libraries of Alabort-i-Medina et al. (2014), Pedregosa et al. (2011). The source code as well as the list of errors for the top methods will be released for the research community in the link https://github.com/grigorisg9gr/deformable_tracking_review_ijcv2016.

4.2.1 Landmark Localisation Training

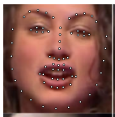
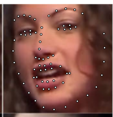
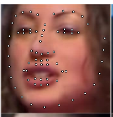
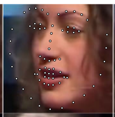
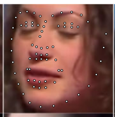
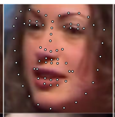
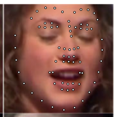
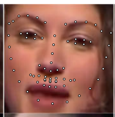
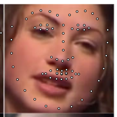
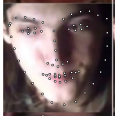
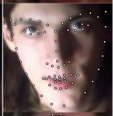
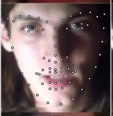
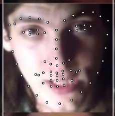
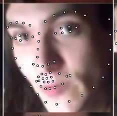
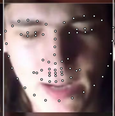
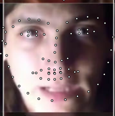
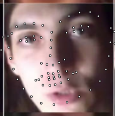
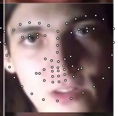
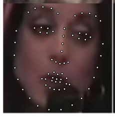
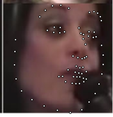
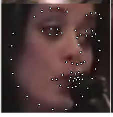
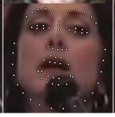
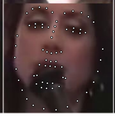
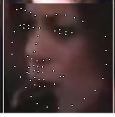
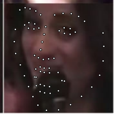
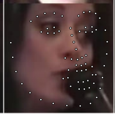
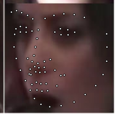
All the landmark localisation methods were trained with respect to the 68 facial points mark-up employed by Sagonas et al. (2013a, 2015) in 300W, while the rest of the parameters were determined via cross-validation. Again, this validation set consisted of frames from the 300 VW trainset, as well as 60 privately collected images with challenging poses. All of the discriminative landmark localisation methods (SDM, ERT, CFSS) were trained from images in the public datasets of IBUG (Sagonas et al. 2013a), HELEN (Le et al. 2012), AFW (Zhu and Ramanan 2012) and LFPW (Belhumeur et al. 2013). The generative AAM was trained on less data, since generative methods do not benefit as strongly from large training datasets. The training data used for the AAM was the recently released 300 images from the 600W dataset (Sagonas et al. 2015), 500 challenging images from LFPW (Belhumeur et al. 2013) and the 135 images of the IBUG dataset (Sagonas et al. 2013a).

Discriminative landmark localisation methods are tightly coupled with the initialisation statistics, as they learn to model a given variance of initialisations. Therefore, it is necessary to re-train each discriminative method for each face detection method employed. This allows the landmark localisation methods to correctly model the large amount of variance present between detectors. On aggregate 5 different detector and landmark localisation models are trained. One for each detector and landmark localisation pair (totalling 4) and a single model trained using a validation set that estimates the variance of the ground truth bounding box throughout the sequences. This model is used for all trackers.

4.2.2 Quantitative Metrics

The errors reported for all the following experiments are with respect to the landmark localisation error. The error metric employed is the mean Euclidean distance of the 68 points, normalised by the diagonal of the ground truth bounding box ($\sqrt{width^2 + height^2}$). This metric was chosen as it is robust to changes in head pose which are frequent within the 300 VW sequences. The graphs that are shown are cumulative error distribution (CED) plots that provide the proportion of images less than or equal to a particular error. We also provide summary tables with respect to the Area Under the Curve (AUC) of the CED plots, considered up to a maximum error. Errors above this maximum threshold, which is fixed to 0.08, are considered failures to accurately localise the facial landmarks. Therefore, we also report the failure

Table 5 Exemplar deformable tracking results that are indicative of the fitting quality that corresponds to each error value for all video categories

Category	Error								
	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
1									
2									
3									

The area under the curve (AUC) and failure rate for all the experiments are computed based on the Cumulative error distributions (CED) limited at maximum error of 0.08

rate, as a percentage, which marks the proportion of images that are not considered within the CED plots. Table 5 shows some indicative examples of the deformable fitting quality that corresponds to each error value for all video categories. When ranking methods, we consider the AUC as the primary statistic and only resort to considering the failure rate in cases where there is little distinction between methods' AUC values.

The indicative speed metric (times) reported in the outcomes is measured on 100 frames of a single video with 640 × 360 resolution. Note that the utlised detectors' performance is highly affected by the resolution. The times were measured in a single machine with a i7 processor, 3.6 GHz, all in CPU mode, with 8GB RAM and report the time in seconds. The implementations were not optimised to minimise the computational complexity, i.e. the public implementations in C/C++ have a considerable advantage.

4.3 Experiment 1: Detection and Landmark Localisation

In this experiment, we validate the most frequently used facial deformable tracking strategy, i.e. performing face detection followed by landmark localisation *on each frame independently*. If a detector fails to return a frame, that frame is considered as having infinite error and thus will appear as part of the failures in Table 6. Note that the AUC is robust to the use of infinite errors. In frames where multiple bounding boxes are returned, the box with the highest confidence is kept, limiting the results of the detectors to a single bounding box per image. A high level diagram explaining the detection procedure for this experiment is given by Fig. 1.

Specifically, in this experiment we consider the 8 face detectors of Table 1 (DPM, HR-TF, MTCNN, NPD, SS-DPM, HOG + SVM, VJ, VPHR) with the 4 landmark localisation techniques of Table 3 (AAM, CFSS, ERT, SDM), for a total of 32 results. The results of the experiment are given in Table 6 and Fig. 3. The results indicate that the AAM performs poorly as it achieves the lowest performance across all face detectors. The discriminative CFSS and ERT landmark localisation methods consistently outperform SDM. From the detectors point of view, it seems that the strongly supervised DPM (SS-DPM) is the worst and provides the highest failure rates. On the other hand, the weakly supervised DPM (DPM) outperforms the rest of the detectors in the first two categories in terms of both accuracy (i.e. AUC) and robustness (i.e. Failure Rate), while in the third one, the deep detector of Zhang et al. (2016) outperforms marginally DPM. In all three categories the state-of-the-art deep networks fetch top results, however they do not seem to be consistently better than DPM or VPHR of Kumar et al. (2015). The detailed graphs per method (32 methods in total), as well as a video with the results of the top five methods (see footnote 5) are deferred to the supplementary material.

4.4 Experiment 2: Detection and Landmark Localisation with Reinitialisation

Complementing the experiments of Sect. 4.3, the same set-up was utilised to study the effect of missed frames by assuming a first order Markov dependency. If the detector does not return a bounding box in a frame, the bounding box of the previous frame is used as a successful detection for the missing frame. This procedure is depicted in Fig. 4. Given that the frame rate of the input videos is adequately high (over 20 fps), this assumption is a reasonable one. The results of this experiment are summarised in Table 7 and in Fig. 5. As expected, the ranking of the methods is almost identical as the previous experiment of Sect. 4.3, with the minor differences emerging from the threshold of the different detectors.

Table 6 Results for experiment 1 of Sect. 4.3 (detection + landmark localisation) (Color table online)

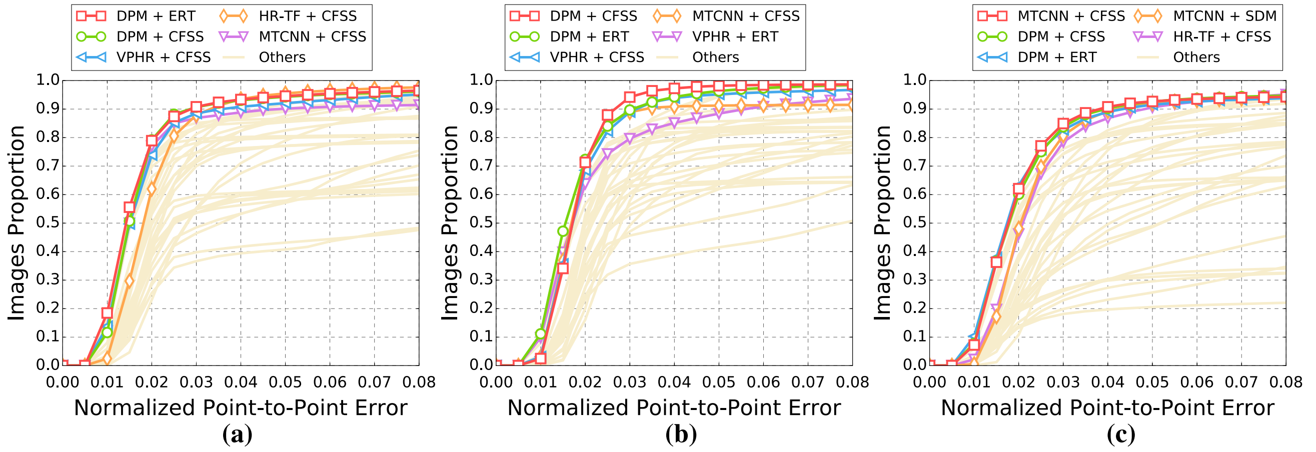
Method		Category 1		Category 2		Category 3		Complexity
Detection	Landmark Localisation	AUC	Failure Rate (%)	AUC	Failure Rate (%)	AUC	Failure Rate (%)	Timing
DPM	AAM	0.447	29.445	0.466	21.158	0.376	33.261	🟧 **2.087**
	CFSS	🟩 **0.764**	🟩 **3.789**	🟥 **0.767**	🟥 **1.363**	🟩 **0.717**	🟩 **5.259**	
	ERT	🟥 **0.772**	🟥 **3.493**	🟩 **0.765**	🟩 **1.558**	🟦 **0.714**	🟦 **6.100**	
	SDM	0.673	3.800	0.646	1.369	0.585	5.880	
HR-TF	AAM	0.468	32.754	0.538	25.240	0.451	29.833	-
	CFSS	🟧 **0.735**	🟧 **2.363**	0.646	13.735	0.677	4.793	
	ERT	0.571	8.898	0.509	18.335	0.538	12.128	
	SDM	0.654	5.499	0.592	14.170	0.612	6.371	
MTCNN	AAM	0.323	51.474	0.406	36.283	0.203	65.005	3.204
	CFSS	0.732	8.553	🟧 **0.722**	🟧 **8.524**	🟥 **0.720**	🟥 **5.685**	
	ERT	0.630	12.299	0.614	10.167	0.636	8.040	
	SDM	0.690	8.203	0.674	8.567	🟧 **0.684**	🟧 **5.772**	
NPD	AAM	0.337	52.230	0.320	48.941	0.263	54.173	🟦 **0.203**
	CFSS	0.492	38.135	0.507	35.571	0.491	34.052	
	ERT	0.461	38.781	0.463	35.787	0.461	34.746	
	SDM	0.451	39.769	0.471	35.754	0.455	34.839	
SS-DPM	AAM	0.474	37.473	0.502	33.807	0.161	77.932	12.400
	CFSS	0.609	21.773	0.566	24.261	0.244	65.926	
	ERT	0.635	21.445	0.608	21.638	0.243	67.407	
	SDM	0.582	21.225	0.537	21.748	0.217	67.602	
SVM+HOG	AAM	0.493	25.891	0.487	22.414	0.380	36.728	🟥 **0.038**
	CFSS	0.707	12.953	0.663	16.318	0.579	21.422	
	ERT	0.705	13.285	0.653	16.500	0.570	22.303	
	SDM	0.654	13.252	0.619	16.312	0.480	21.367	
VJ	AAM	0.453	24.277	0.532	19.500	0.413	25.640	🟩 **0.052**
	CFSS	0.660	18.986	0.651	17.805	0.641	15.061	
	ERT	0.658	19.292	0.646	17.839	0.653	14.942	
	SDM	0.524	19.249	0.548	17.769	0.505	15.347	
VPHR	AAM	0.463	34.436	0.636	12.737	0.519	23.065	30.200
	CFSS	🟦 **0.747**	🟦 **4.860**	🟦 **0.743**	🟦 **3.255**	0.652	11.287	
	ERT	0.725	6.834	0.700	6.328	0.624	13.490	
	SDM	0.661	7.367	0.655	6.239	0.549	15.206	

Colouring denotes the methods' performance ranking per category: 🟥 first 🟩 second 🟦 third 🟧 fourth

The area under the curve (AUC) and Failure Rate are reported. The top four performing curves are highlighted for each video category. The current implementation of HR-TF cannot be executed to CPU mode, thus it would be unfair for the rest of the timing comparisons to include its GPU performance

For instance, the SVM + HOG that has a high threshold, i.e. in the previous experiment it 'missed' several challenging frames, can benefit further from the Markov dependency, while the VPHR one has exactly the same statistics as it returned a detection in every single frame in the previous experiment.

In order to better investigate the effect of this reinitialisation scheme, we also provide Fig. 6 that directly shows the improvement. Specifically, we plot the CED curves with and without the reinitialisation strategy for 3 top performing methods, as well as the 3 techniques for which the highest improvement is achieved. It becomes evident that the top performing methods from Sect. 4.3 do not benefit from reinitialisation, since the improvement is marginal. This is explained by the fact that these methods already achieve a very high true positive rate. The largest difference is observed for methods that utilise AAM. As shown by Antonakos et al. (2015b), AAMs are very sensitive to initialisation, due to the nature of Gauss-Newton optimisation. Additionally, note that we have not attempted to apply any kind of greedy approach for improving the detectors' bounding boxes in order to provide a better AAM initialisation. Since the initialisation of a



(a) **(b)** **(c)**

Fig. 3 Results for experiment 1 of Sect. 4.3 (detection + landmark localisation). The top 5 performing curves are highlighted in each legend. Please see Table 6 for a full summary

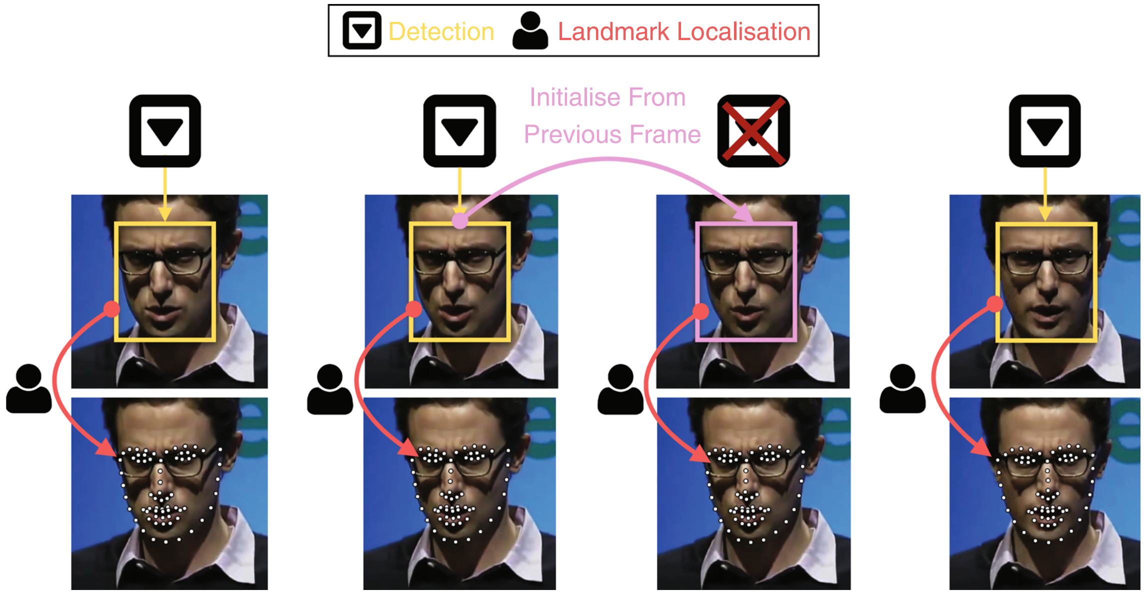


Fig. 4 This figure gives a diagram of the reinitialisation scheme proposed in Sect. 4.4. Specifically, in case the face detector does not return a bounding box for a frame, the bounding box of the previous frame is used as a successful detection for the missing frame

frame with failed detection is achieved by the bounding box of the previous frame's landmarks, it is highly likely that its area will be well constrained to include only the facial parts and not the forehead or background. This kind of initialisation is very beneficial for AAMs, which justifies the large improvements that are shown in Fig. 6. For the graphs that correspond to all 32 methods, please refer to the supplementary material.

4.5 Experiment 3: Model-free Tracking and Landmark Localisation

In this section, we provide, to the best of our knowledge, the first detailed analysis of the performance of model free trackers for tracking "in-the-wild" facial sequences. For this reason, we have considered a large number of trackers in order to attempt to give a balanced overview of the performance of modern model trackers for deformable face alignment. The 27 trackers considered in this section are summarised in Table 2. To initialise all trackers, the tightest possible bounding box of the ground truth facial landmarks is provided as the initial tracker state. We also include a baseline method, which appears in results Table 8, referred to as PREV, which is defined as applying the landmark localisation methods initialised from the bounding box of the result in the previous frame. Obviously this scheme is highly sensitive to drifting and therefore we have included it as a basic baseline that does not include any model free tracking. A

Table 7 Results for experiment 2 of Sect. 4.4 (detection + landmark localisation + initialisation from previous frame) (Color table online)

Method		Category 1		Category 2		Category 3	
Detection	Landmark Localisation	AUC	Failure Rate (%)	AUC	Failure Rate (%)	AUC	Failure Rate (%)
DPM	AAM	0.572	18.840	0.621	10.617	0.493	21.711
	CFSS	🟩 **0.765**	🟩 **3.415**	🟥 **0.769**	🟥 **0.815**	🟩 **0.720**	🟩 **4.786**
	ERT	🟥 **0.773**	🟥 **3.221**	🟩 **0.767**	🟩 **1.156**	🟦 **0.716**	🟦 **5.620**
	SDM	0.674	3.727	0.654	1.129	0.579	6.006
HR-TF	AAM	0.468	32.754	0.538	25.240	0.451	29.833
	CFSS	0.735	2.363	0.653	12.844	0.677	4.786
	ERT	0.571	8.877	0.513	17.519	0.538	12.132
	SDM	0.654	5.483	0.598	13.288	0.612	6.368
MTCNN	AAM	0.323	51.474	0.406	36.283	0.203	65.005
	CFSS	0.748	6.055	🟦 **0.760**	🟦 **2.717**	🟥 **0.726**	🟥 **4.388**
	ERT	0.639	10.429	0.633	5.503	0.639	7.220
	SDM	0.705	5.747	0.711	2.604	🟧 **0.689**	🟧 **4.674**
NPD	AAM	0.337	52.227	0.320	48.941	0.264	54.141
	CFSS	0.494	37.742	0.511	34.841	0.499	32.625
	ERT	0.463	38.436	0.467	35.036	0.466	33.521
	SDM	0.452	39.416	0.476	34.984	0.462	33.467
SS-DPM	AAM	0.507	32.867	0.526	28.781	0.175	75.646
	CFSS	0.609	21.734	0.576	22.070	0.248	65.421
	ERT	0.636	21.397	0.622	18.459	0.246	66.905
	SDM	0.594	21.306	0.569	18.444	0.227	67.653
SVM+HOG	AAM	0.627	13.770	0.643	11.210	0.526	20.215
	CFSS	🟦 **0.759**	🟦 **5.009**	🟧 **0.747**	🟧 **4.186**	0.632	12.179
	ERT	🟧 **0.750**	🟧 **6.002**	0.717	6.428	0.615	13.963
	SDM	0.685	6.218	0.676	6.325	0.522	13.234
VJ	AAM	0.570	18.339	0.593	15.612	0.546	16.831
	CFSS	0.685	14.945	0.686	12.619	0.660	11.612
	ERT	0.679	15.783	0.675	12.862	0.672	11.543
	SDM	0.536	16.452	0.573	13.175	0.530	12.779
VPHR	AAM	0.482	28.893	0.636	12.737	0.519	23.065
	CFSS	0.747	4.860	0.743	3.255	0.652	11.287
	ERT	0.725	6.834	0.700	6.328	0.624	13.490
	SDM	0.661	7.367	0.655	6.239	0.549	15.206

Colouring denotes the methods' performance ranking per category: 🟥 first 🟩 second 🟦 third 🟧 fourth

The area under the curve (AUC) and failure rate are reported. The top four performing curves are highlighted for each video category

high level diagram explaining the detection procedure for this experiment is given by Fig. 1.

Specifically, in this experiment we consider the 27 model free trackers of Table 2, plus the PREV baseline, with the 4 landmark localisation techniques of Table 3 (AAM, CFSS, ERT, SDM), for a total of 112 results. The results of the experiment are given in Table 8 and Fig. 7. Please see the supplementary material for full statistics.

By inspecting the results, we can firstly notice that most generative trackers perform poorly (i.e. ORIA, DF, FCT, IVT), except LRST which achieves decent performance for the most challenging video category.The discriminative approaches of SRDCF and SPOT are consistently performing very well, however the trackers employing deep neural networks fetch the most accurate outcomes, consistent with the latest VOT competition outcomes. Additionally, similar to the face detection experiments, the combination of all trackers with CFSS returns the best result, whereas AAM constantly demonstrates the poorest performance. Finally, it becomes evident that a straightforward application of the simplistic baseline approach (PREV) is not suitable for deformable tracking, even though it is surprisingly outperforming some model free trackers, such as DF, ORIA and FCT. For the curves that correspond to all 112 methods as well as a video with the tracking result of the top five methods (see footnote 6), please refer to the supplementary material.

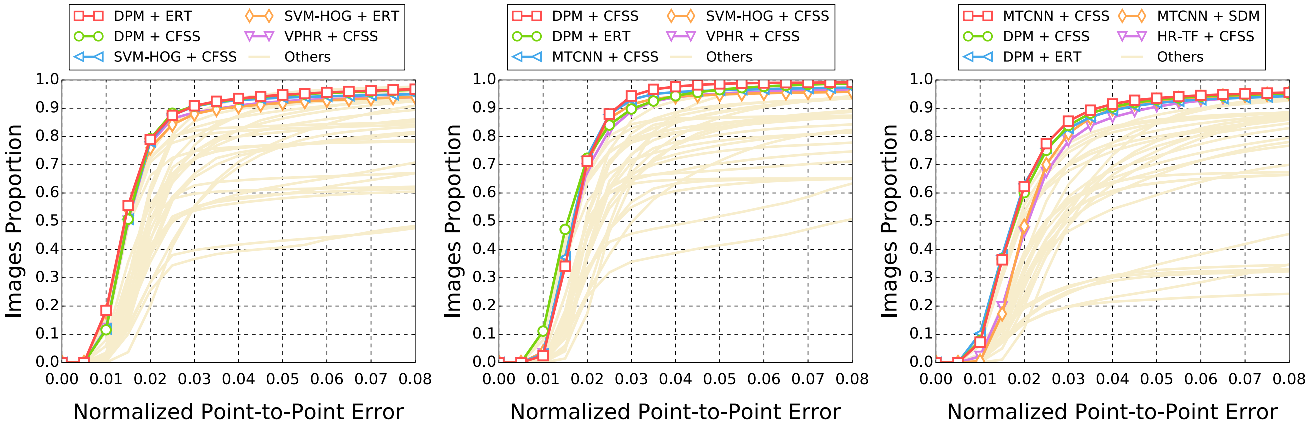
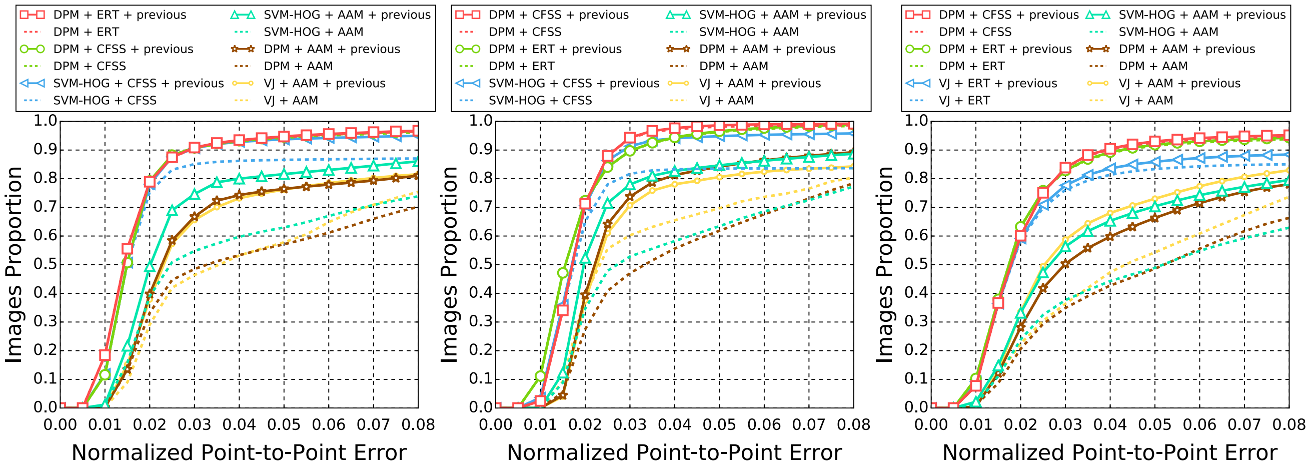


Fig. 5 Results for experiment 2 of Sect. 4.4 (detection + landmark localisation + initialisation from previous frame). The top five performing *curves* are highlighted in each legend. Please see Table 7 for a full summary



(a) (b) (c)

Fig. 6 Results for experiment 2 of Sect. 4.4 (detection + landmark localisation + initialisation from previous frame). These results show the effect of initialisation from the previous frame, in comparison to missing detections. The top three performing results are given in *red*, *green* and *blue*, respectively, and the top three most improved are given in *cyan*, *yellow* and *brown*, respectively. The *dashed lines* represent the results before the reinitialisation strategy is applied, *solid lines* are after (Color figure online)

4.6 Experiment 4: Failure Checking and Tracking Reinitialisation

Complementing the experiments of Sect. 4.5, we investigate the improvement in performance of performing failure checking during tracking. Here we define failure checking as the process of determining whether or not the currently tracked object is a face. Given that we have prior knowledge of the class of object we are tracking, namely faces, this enables us to train an offline classifier that attempts to determine whether a given input is a face or not. Furthermore, since we are also applying landmark localisation, we can perform a strong classification by using the facial landmarks as position priors when extracting features for the failure checking. To train the failure checking classifier, we perform the following methodology:

1. For all images in the Landmark Localisation training set, extract a fixed sized patch around each of the 68 landmarks and compute HOG (Dalal and Triggs 2005) features for each patch. These patches are the positive training samples.
2. Generate negative training samples by perturbing the ground truth bounding box, extracting fixed size patches and computing HOG.
3. Train an SVM classifier using the positive and negative samples.

Table 8 Results for experiment 3 of Sect. 4.5 (model free tracking + landmark localisation) (Color table online)

Method		Category 1		Category 2		Category 3		Complexity
Rigid Tracking	Landmark Localisation	AUC	Failure Rate (%)	AUC	Failure Rate (%)	AUC	Failure Rate (%)	Timing
PREV	AAM	0.375	50.652	0.465	38.273	0.095	87.734	0
	CFSS	0.545	27.358	0.618	19.865	0.199	72.991	
	ERT	0.340	57.266	0.438	42.011	0.073	89.959	
	SDM	0.497	36.606	0.505	32.843	0.194	74.111	
CAMSHIFT	AAM	0.030	94.900	0.053	88.051	0.023	95.604	**0.012**
	CFSS	0.062	86.809	0.128	72.524	0.079	79.926	
	ERT	0.032	93.046	0.030	90.007	0.026	92.336	
	SDM	0.039	89.794	0.072	81.808	0.031	89.295	
CCOT	AAM	0.561	22.570	0.673	9.905	0.412	34.298	1.500
	CFSS	0.719	7.748	**0.771**	**1.235**	**0.698**	**4.305**	
	ERT	0.667	10.516	0.724	4.043	0.570	10.485	
	SDM	0.654	9.099	0.703	2.458	0.592	9.737	
CMT	AAM	0.574	20.323	0.691	8.424	0.478	26.334	0.038
	CFSS	0.748	2.635	0.758	1.871	0.595	16.506	
	ERT	0.653	6.950	0.716	2.847	0.498	21.136	
	SDM	0.669	3.808	0.706	2.184	0.529	18.427	
DF	AAM	0.270	60.722	0.290	57.404	0.224	67.165	0.185
	CFSS	0.467	38.756	0.460	35.465	0.348	51.761	
	ERT	0.337	48.838	0.344	46.094	0.246	59.526	
	SDM	0.358	47.286	0.365	43.672	0.275	57.901	
DLSSVM	AAM	0.566	21.800	0.671	8.052	0.403	32.896	0.126
	CFSS	**0.762**	**2.503**	0.748	0.459	0.612	15.256	
	ERT	0.680	6.075	0.640	4.557	0.456	23.011	
	SDM	0.694	3.860	0.659	1.609	0.512	20.291	
DSST	AAM	0.510	28.620	0.675	8.442	0.246	59.761	**0.014**
	CFSS	0.670	13.018	0.764	0.605	0.380	44.205	
	ERT	0.549	17.341	0.686	2.434	0.286	48.893	
	SDM	0.552	14.509	0.686	1.558	0.304	46.433	
FCT	AAM	0.341	51.592	0.549	20.288	0.148	76.888	**0.009**
	CFSS	0.527	29.347	0.706	9.409	0.319	53.043	
	ERT	0.384	40.603	0.619	11.989	0.187	65.215	
	SDM	0.418	38.522	0.627	12.524	0.203	63.803	
HDT	AAM	0.422	40.148	0.558	23.500	0.268	56.182	1.047
	CFSS	0.631	19.268	0.684	9.6100	0.534	26.399	
	ERT	0.491	27.749	0.615	15.983	0.343	37.566	
	SDM	0.525	26.141	0.603	14.280	0.399	35.811	
IVT	AAM	0.429	40.724	0.424	42.699	0.245	61.675	0.020
	CFSS	0.580	28.005	0.533	28.225	0.423	42.244	
	ERT	0.507	31.802	0.477	32.773	0.329	47.033	
	SDM	0.517	30.971	0.464	33.706	0.348	45.664	
KCF	AAM	0.550	25.025	0.672	8.731	0.376	39.221	**0.011**
	CFSS	0.693	11.221	0.741	2.847	0.554	16.889	
	ERT	0.642	13.318	0.716	3.714	0.438	24.838	
	SDM	0.626	12.119	0.694	3.069	0.444	22.686	
LCT	AAM	0.534	26.336	0.670	9.248	0.435	31.694	0.172
	CFSS	0.706	10.527	0.770	0.627	0.644	12.88	
	ERT	0.660	12.903	0.731	1.898	0.531	16.123	
	SDM	0.650	12.025	0.710	2.172	0.568	14.761	
LRST	AAM	0.537	26.997	0.633	13.419	0.426	32.878	1.738
	CFSS	0.704	10.873	0.759	1.600	0.649	13.526	
	ERT	0.629	13.191	0.698	4.429	0.531	16.712	
	SDM	0.643	12.730	0.696	4.040	0.580	15.249	
MDNET	AAM	0.579	19.944	0.649	9.354	0.500	24.893	3.101
	CFSS	**0.780**	**1.789**	**0.780**	**0.383**	**0.706**	**7.520**	
	ERT	**0.758**	**2.390**	0.762	0.812	0.632	9.972	
	SDM	0.734	2.137	0.732	1.238	0.653	8.647	

Table 8 continued

Method		Category 1		Category 2		Category 3		Complexity
Rigid Tracking	Landmark Localisation	AUC	Failure Rate (%)	AUC	Failure Rate (%)	AUC	Failure Rate (%)	Timing
MEEM	AAM	0.493	29.022	0.605	12.299	0.370	41.680	0.102
	CFSS	**0.761**	**3.534**	**0.775**	**0.420**	**0.662**	**11.236**	
	ERT	0.647	8.874	0.728	0.989	0.545	13.223	
	SDM	0.666	7.283	0.717	0.998	0.598	13.071	
MIL	AAM	0.445	32.327	0.544	21.654	0.185	67.093	0.075
	CFSS	0.683	11.420	0.710	4.128	0.380	45.910	
	ERT	0.536	16.881	0.603	10.413	0.237	57.771	
	SDM	0.589	14.693	0.626	8.746	0.268	56.023	
ORIA	AAM	0.364	48.718	0.566	21.17	0.128	77.014	0.076
	CFSS	0.501	34.015	0.665	10.617	0.273	60.909	
	ERT	0.436	38.251	0.640	12.491	0.227	61.343	
	SDM	0.395	43.986	0.634	12.144	0.188	66.970	
PF	AAM	0.297	54.275	0.428	34.680	0.108	78.217	0.088
	CFSS	0.546	29.095	0.616	15.296	0.415	38.101	
	ERT	0.399	37.648	0.457	26.530	0.240	50.804	
	SDM	0.445	35.104	0.504	23.351	0.294	48.817	
RPT	AAM	0.477	32.206	0.617	12.181	0.379	39.640	0.348
	CFSS	0.725	5.751	0.768	0.271	0.627	13.324	
	ERT	0.587	12.897	0.709	2.388	0.506	18.698	
	SDM	0.620	9.191	0.708	0.925	0.538	17.539	
SIAM-OXF	AAM	0.498	31.921	0.648	11.879	0.500	25.496	0.220
	CFSS	0.714	8.200	0.740	2.333	0.653	10.420	
	ERT	0.648	12.077	0.688	6.239	0.564	15.538	
	SDM	0.633	11.398	0.671	5.540	0.567	13.775	
SPOT	AAM	0.535	25.227	0.680	7.058	0.253	57.121	0.154
	CFSS	**0.769**	**2.330**	**0.774**	**0.435**	0.546	27.414	
	ERT	0.638	6.809	0.728	1.095	0.411	30.458	
	SDM	0.679	3.244	0.715	0.532	0.472	28.562	
SPT	AAM	0.141	77.871	0.105	81.087	0.051	90.949	1.438
	CFSS	0.267	64.039	0.216	67.550	0.157	76.707	
	ERT	0.178	70.526	0.123	76.454	0.084	83.288	
	SDM	0.202	68.807	0.144	74.203	0.102	82.150	
SRDCF	AAM	0.545	26.056	0.675	7.824	0.437	31.827	0.206
	CFSS	0.731	6.810	**0.779**	**0.155**	**0.687**	**8.145**	
	ERT	0.636	11.251	0.743	0.980	0.544	11.666	
	SDM	0.650	7.929	0.726	0.435	0.587	10.788	
STAPLE	AAM	0.503	28.187	0.673	8.049	0.389	35.118	0.035
	CFSS	0.686	8.048	0.767	0.541	**0.656**	**6.787**	
	ERT	0.573	15.567	0.702	2.653	0.486	14.487	
	SDM	0.587	12.706	0.692	2.367	0.533	11.666	
STCL	AAM	0.075	87.172	0.124	80.074	0.035	92.878	**0.015**
	CFSS	0.163	73.950	0.176	72.289	0.095	84.473	
	ERT	0.086	82.045	0.121	77.580	0.050	88.695	
	SDM	0.094	80.686	0.121	76.740	0.049	89.385	
STRUCK	AAM	0.543	25.041	0.648	13.282	0.360	42.496	0.028
	CFSS	0.728	7.741	0.741	4.411	0.585	21.050	
	ERT	0.596	11.148	0.685	5.528	0.430	27.139	
	SDM	0.643	8.866	0.681	4.965	0.488	25.156	
TGPR	AAM	0.504	26.583	0.634	13.148	0.361	42.843	1.353
	CFSS	0.721	6.866	0.757	1.901	0.625	13.714	
	ERT	0.606	10.166	0.689	3.854	0.454	21.866	
	SDM	0.623	7.892	0.687	4.067	0.488	20.143	
TLD	AAM	0.373	42.618	0.507	18.837	0.269	55.885	0.066
	CFSS	0.622	14.940	0.678	7.502	0.469	29.592	
	ERT	0.410	30.337	0.544	14.952	0.302	38.877	
	SDM	0.456	25.006	0.564	11.676	0.333	37.440	

Colouring denotes the methods' performance ranking per category: 🟥 first 🟩 second 🟦 third 🟧 fourth 🟪 fifth

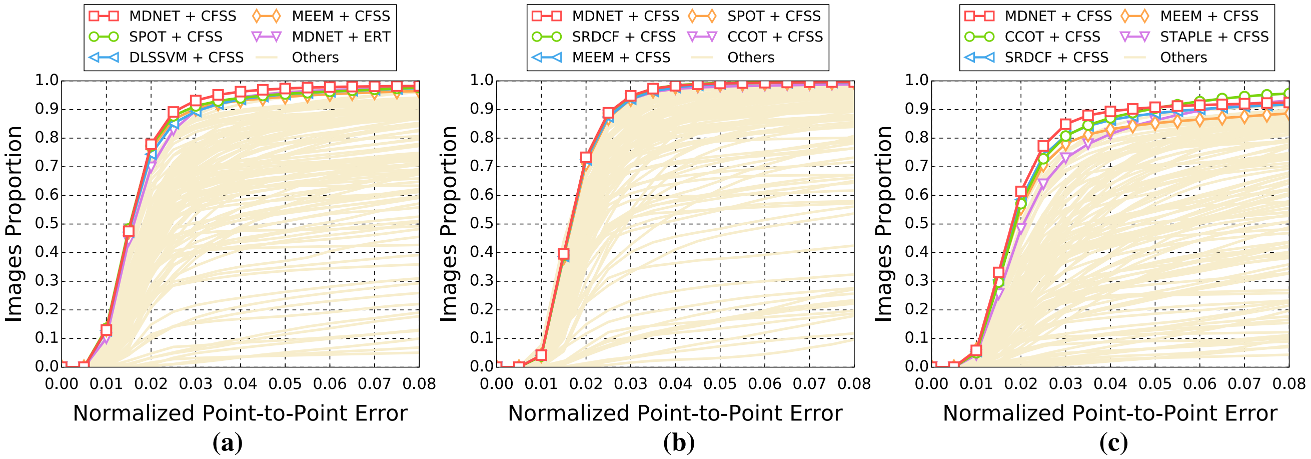


Fig. 7 Results for experiment 3 of Sect. 4.5 (model free tracking + landmark localisation). The top five performing *curves* are highlighted in each legend. Please see Table 8 for a full summary

For the experiments in this section, we use a fixed patch size of 18 × 18 pixels, with 100 negative patches sampled for each positive patch. The failure checking classification threshold is chosen via cross-validation on two sequences from the 300 VW training videos. Any hyper-parameters of the SVM are also trained using these two validation videos.

Given the failure detector, our restart procedure, is as follows:

- Classify the current frame to determine if the tracking has failed. If a failure is verified, perform a restart, otherwise continue.
- Following the convention of the VOT challenges by Kristan et al. (2013, 2014, 2015), we attempt to reduce the probability that poor trackers will overly rely on the output of the failure detection system. In the worst case, a very poor tracker would fail on most frames and thus the accuracy of the detector would be validated rather than the tracker itself. Therefore, when a failure is identified, the tracker is allowed to continue for 10 more frames. The results from the drifting tracker are used in these 10 frames in order reduce the affect of the detector. The tracker is then reinitialised at the frame it was first detected as failing at. The next 10 frames, as previously described, already have results computed and therefore no landmark localisation or failure checking is performed in these frames. At the 11th frame, the tracker continues as normal, with landmark localisation and failure checking.
- In the unlikely event that the detector fails to detect the face, the previous frame is used as described in Sect. 4.4.

The diagram given in Fig. 8 gives a pictorial representation of this scheme.

The results of this experiment are given in Table 9 and Fig. 9. In contrast to Sect. 4.5, we only perform the experiments on a subset of the total trackers using CFSS. We use 3 among the top performing trackers (SRDCF, RPT, SPOT) as well as FCT which had mediocre performance in Sect. 4.5. The results indicate that SRDCF is the best model free tracking methodology for the task.

In order to better investigate the effect of this failure checking scheme, we also provide Fig. 6 which shows the differences between the initial tracking results of Sect. 4.5 and the results after applying failure detection. The performance of top trackers (i.e. SRDCF, SPOT, RPT) does not improve much, which is expected since they are already able to return a robust tracking result. However, FCT benefits from the failure checking process, which apparently minimises its drifting issues.

4.7 Experiment 5: Kalman Smoothing

In this section, we report the effect of performing Kalman Smoothing (Kalman 1960) on the results of the detectors of Sect. 4.3 and the trackers of Sect. 4.5. This experiment is designed to highlight the stability of the current landmark localisation methods with respect to noisy movement between frames (or jittering as it often known). However, when attempting to smooth the trajectories of the tracked bounding boxes themselves, we found an extremely negative effect on the results. Therefore, to remove jitter from the results we perform Kalman smoothing on the landmarks themselves. To robustly smooth the landmark trajectories, a generic facial shape model is constructed in a similar manner as described in the AAM literature by Cootes et al. (2001). Specifically, given the sparse shape of the face consisting of n landmark points, we denote the coordinates of the i-th landmark point within the Cartesian space of the

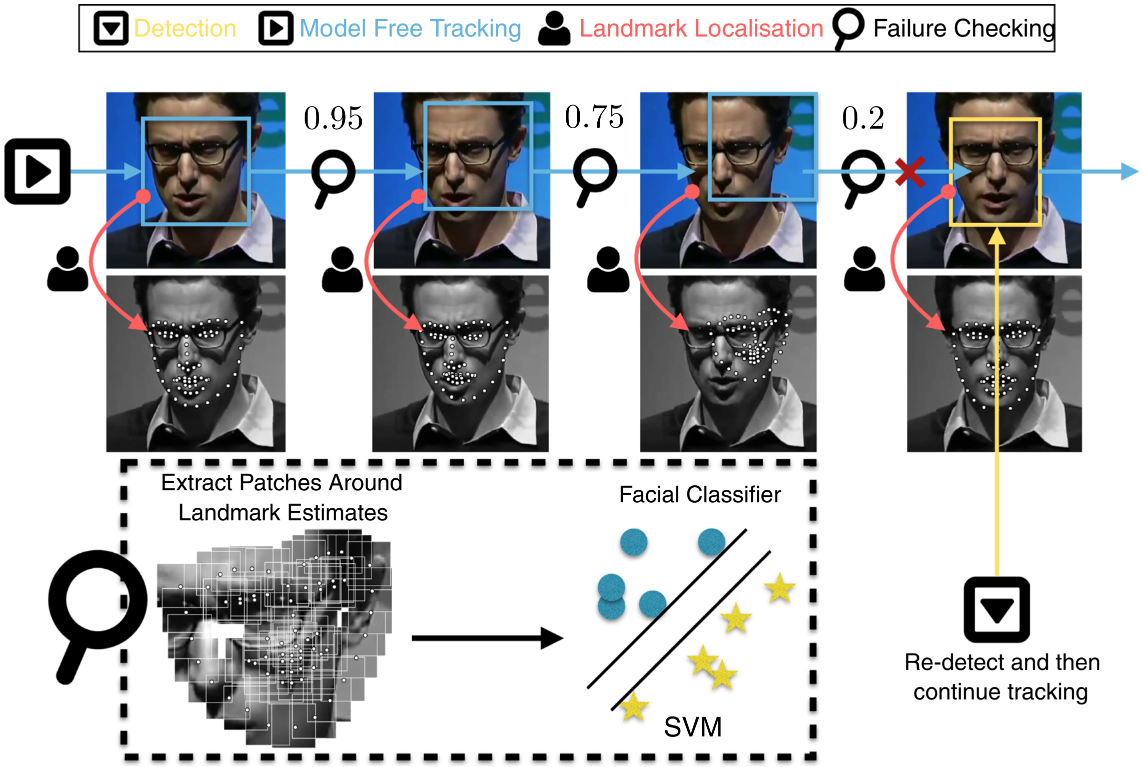


Fig. 8 This figure gives a diagram of the reinitialisation scheme proposed in Sect. 4.6 for tracking with failure detection. For all frames after the first, the result of the current landmark localisation is used to decide whether or not a face is still being tracked. If the classification fails, a re-detection is performed and the tracker is reinitialised with the bounding box returned by the detector

Table 9 Results for experiment 4 of Sect. 4.6 (model free tracking + landmark localisation + failure checking) (Color table online)

Method		Category 1		Category 2		Category 3	
Rigid Tracking	Landmark Localisation	AUC	Failure Rate (%)	AUC	Failure Rate (%)	AUC	Failure Rate (%)
FCT	CFSS	**0.693**	**13.414**	**0.763**	**1.661**	0.516	32.376
RPT		**0.745**	**6.239**	**0.769**	**0.697**	**0.704**	**6.108**
SPOT		0.688	13.342	0.751	2.896	**0.570**	**22.913**
SRDCF		**0.748**	**5.999**	**0.772**	**0.505**	**0.698**	**6.657**

Colouring denotes the methods' performance ranking per category: 🟥 first 🟩 second 🟦 third

The area under the curve (AUC) and failure rate are reported. The top 3 performing curves are highlighted for each video category

image $\mathbf{I}$ as $\mathbf{x}_i = [x_i, y_i]^T$. Then a *shape instance* of the face is given by the $2n \times 1$ vector $\mathbf{s} = \left[\mathbf{x}_1^T, \ldots, \mathbf{x}_n^T\right]^T = [x_1, y_1, \ldots, x_n, y_n]^T$. Given a set of N such shape samples $\{\mathbf{s}^1, \ldots, \mathbf{s}^N\}$, a parametric statistical subspace of the object's shape variance can be retrieved by first applying Generalised Procrustes Analysis on the shapes to normalise them with respect to the global similarity transform (i.e., scale, in-plane rotation and translation) and then using Principal Component Analysis (PCA). The resulting *shape model*, denoted as $\{\mathbf{U}_s, \bar{\mathbf{s}}\}$, consists of the orthonormal basis $\mathbf{U}_s \in \mathbb{R}^{2n \times n_s}$ with n_s eigenvectors and the mean shape vector $\bar{\mathbf{s}} \in \mathbb{R}^{2n}$. This parametric model can be used to generate new shape instances as $\mathbf{s}(\mathbf{p}) = \bar{\mathbf{s}} + \mathbf{U}_s\mathbf{p}$ where $\mathbf{p} = [p_1, \ldots, p_{n_s}]^T$ is the $n_s \times 1$ vector of *shape parameters* that control the linear combination of the eigenvectors. The Kalman smoothing is thus learnt via Expectation-Maximisation (EM) for the parameters $\mathbf{p}$ of each shape within a sequence (Fig. 10).

The results of this experiment are given in Table 10 and Fig. 11. These experiments also provide a direct comparison between the best detection and model free tracking based

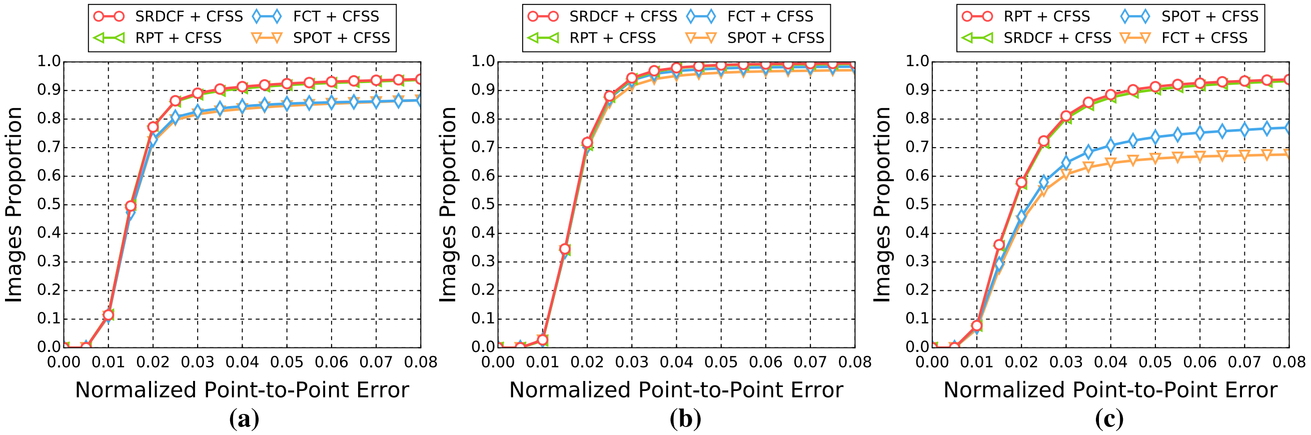


Fig. 9 Results for experiment 4 of Sect. 4.6 (model free tracking + landmark localisation + failure checking). The top five performing *curves* are highlighted in each legend. Please see Table 9 for a full summary

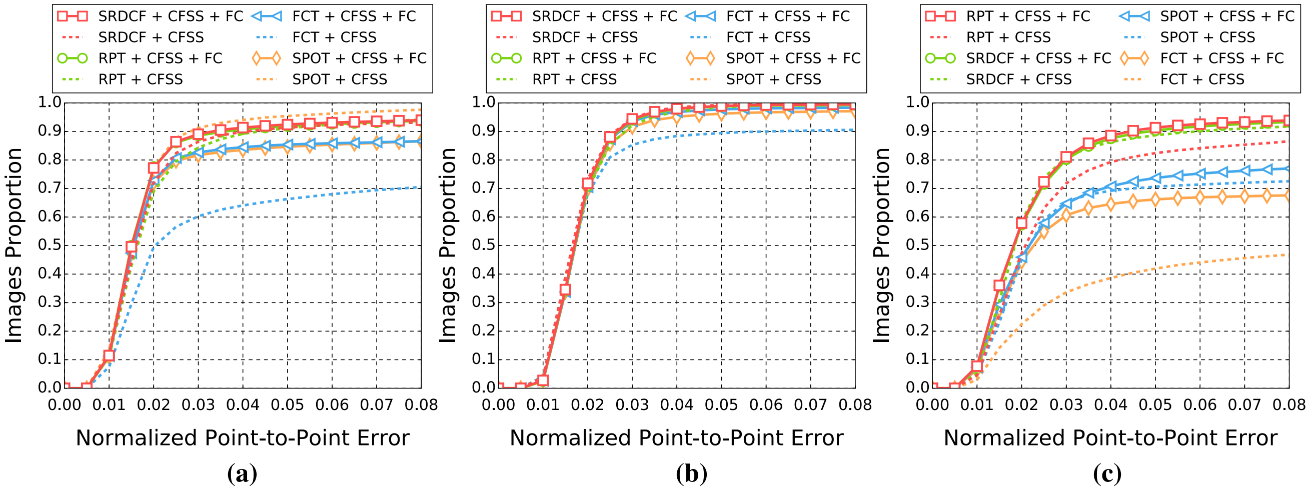


Fig. 10 Results for experiment 4 of Sect. 4.6 (model free tracking + landmark localisation + failure checking). These results show the effect of the failure checking, in comparison to only tracking. The results are coloured by their performance *red*, *green*, *blue* and *orange*, respectively. The *dashed lines* represent the results before the reinitialisation strategy is applied, *solid lines* are after (Color figure online)

techniques. In categories 1 and 2 the Kalman smoothing applied to the model free trackers followed by the discriminative landmark localisation methods of ERT or CFSS score better, with the trackers MDNET and SRDCF being the top performers. In category 3 the DPM and the deep tracker MTCNN achieve the top performance, because they are less prone to drifting (in comparison to trackers) in the most challenging clips of the dataset.

In order to better investigate the effect of the smoothing, we also provide Fig. 12 which shows the differences between the initial tracking results and the results after applying Kalman smoothing. This comparison is shown for the best methods of Table 10. It becomes obvious that the improvement introduced by Kalman smoothing is consistent, but marginal.

4.8 300 VW Comparison

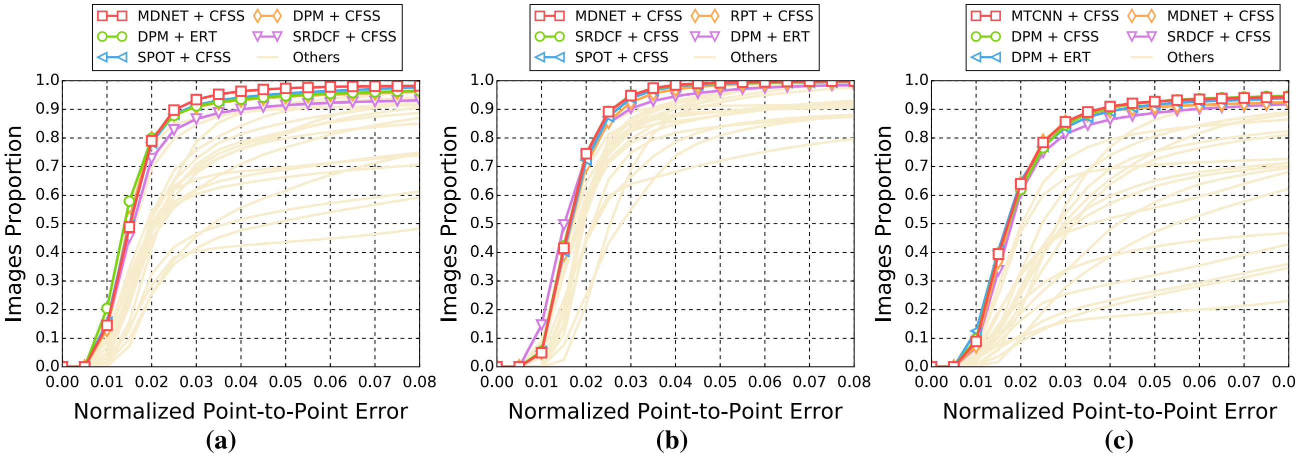
In this section we provide results that compare the best performing methods of the previous Sects. (4.3–4.7) to the participants of the 300 VW challenge by Shen et al. (2015). The challenge had 5 competitors. Rajamanoharan and Cootes (2015) employ a multi-view Constrained Local Model (CLM) with a global shape model and different response maps per pose and explore shape-space clustering strategies to determine the optimal pose-specific CLM. Uricar and Franc (2015) apply a DPM at each frame as well as Kalman smoothing on the face positions. Wu and Ji (2015) utilise a shape augmented regression model, where the regression function is automatically selected based on the facial shape. Xiao et al. (2015) propose a multi-stage regression-based approach that progressively provides ini-

Table 10 Results for experiment 5 of Sect. 4.7 (Kalman Smoothing) (Color table online)

Method		Category 1		Category 2		Category 3	
Detection or Tracking	Landmark Localisation	AUC	Failure Rate (%)	AUC	Failure Rate (%)	AUC	Failure Rate (%)
DPM	CFSS	**0.766**	**3.741**	0.770	1.317	**0.724**	**5.234**
	ERT	**0.777**	**3.442**	0.772	1.509	**0.721**	**6.082**
	SDM	0.678	3.728	0.652	1.354	0.592	5.786
MTCNN	CFSS	0.734	8.507	0.725	8.518	**0.726**	**5.685**
FCT	AAM	0.342	51.503	0.552	20.172	0.149	76.765
	CFSS	0.529	29.283	0.709	9.358	0.320	53.061
	ERT	0.386	40.506	0.623	11.937	0.188	65.121
	SDM	0.419	38.506	0.629	12.515	0.204	63.730
MDNET	CFSS	**0.784**	**1.754**	**0.783**	**0.341**	**0.713**	**7.466**
RPT	CFSS	0.727	5.722	**0.772**	**0.252**	0.632	13.331
	ERT	0.589	12.765	0.713	2.303	0.507	18.687
	SDM	0.622	9.169	0.710	0.888	0.539	17.535
SPOT	AAM	0.536	24.998	0.682	6.957	0.254	56.803
	CFSS	**0.773**	**2.237**	**0.777**	**0.417**	0.551	27.323
	ERT	0.640	6.745	0.731	1.074	0.412	30.296
	SDM	0.681	3.194	0.717	0.508	0.474	28.548
SRDCF	AAM	0.546	25.988	0.676	7.697	0.440	31.499
	CFSS	0.734	6.815	**0.783**	**0.131**	0.693	8.134
	ERT	0.637	11.145	0.746	0.922	0.544	11.572
	SDM	0.652	7.905	0.729	0.414	0.588	10.774
TLD	CFSS	0.624	14.827	0.681	7.477	0.473	29.548
	SDM	0.457	24.965	0.566	11.645	0.335	37.389

Colouring denotes the methods' performance ranking per category: 🟥 first 🟩 second 🟦 third 🟧 fourth

The area under the curve (AUC) and failure rate are reported. The top four performing curves are highlighted for each video category

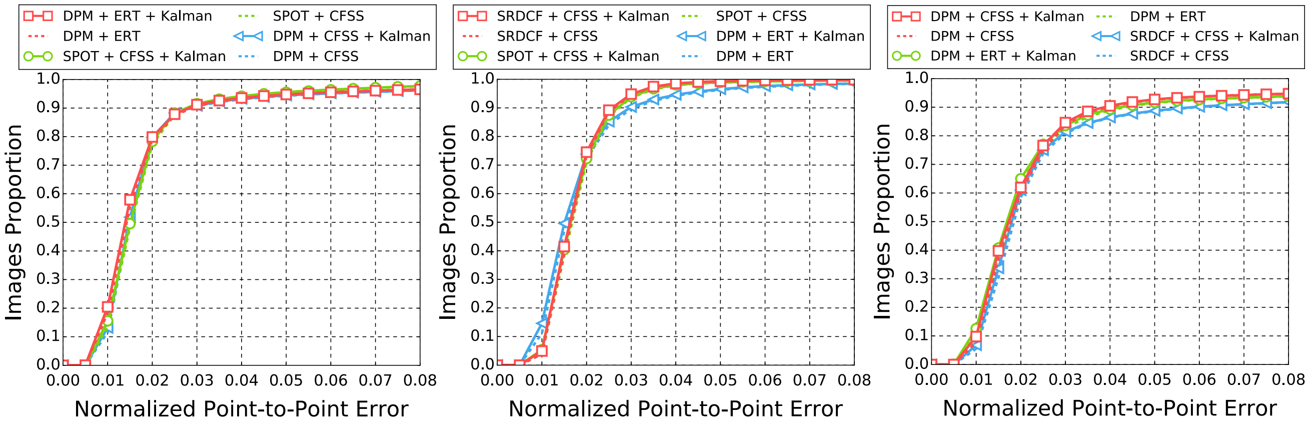


(a) (b) (c)

Fig. 11 Results for experiment 5 of Sect. 4.7 (Kalman Smoothing). The top five performing *curves* are highlighted in each legend. Please see Table 10 for a full summary

tialisations for ambiguous landmarks such as boundary and eyebrows, based on landmarks with semantically strong meaning such as eyes and mouth corners. Finally, Yang et al. (2015a) employ a multi-view spatio-temporal cascade shape regression model along with a novel reinitialisation mechanism.

The results are summarised in Table 11 and Fig. 13. Note that the error metric considered in this paper (as described in



(a) (b) (c)

Fig. 12 Results for experiment 5 of Sect. 4.7 (Kalman Smoothing). These results show the effect of Kalman smoothing on the final landmark localisation results. The top three performing results are given in *red*, *green* and *blue*, respectively, and the top three most improved are given in *cyan*, *yellow* and *brown*, respectively. The *dashed lines* represent the results before the smoothing is applied, *solid lines* are after (Color figure online)

Table 11 Comparison between the best methods of Sects. 4.3–4.7 and the participants of the 300 VW challenge by Shen et al. (2015) (Color table online)

Method	Category 1		Category 2		Category 3	
	AUC	Failure Rate (%)	AUC	Failure Rate (%)	AUC	Failure Rate (%)
DPM + ERT + Kalman	**0.775**	**3.472**	0.770	1.527	**0.719**	**6.111**
DPM + ERT + previous	**0.771**	**3.262**	0.764	1.205	**0.714**	**5.692**
DPM + CFSS + Kalman	**0.764**	**3.784**	0.767	1.326	**0.721**	**5.255**
MDNET + CFSS + Kalman	**0.784**	**1.754**	**0.783**	**0.341**	0.713	7.466
MTCNN + CFSS + Kalman	0.734	8.507	0.725	8.518	**0.726**	**5.685**
MTCNN + CFSS + previous	0.748	6.055	0.760	2.717	**0.726**	**4.388**
SRDCF + CFSS + Kalman	0.732	6.847	**0.780**	**0.131**	0.690	8.206
SRDCF + CFSS	0.729	6.849	**0.777**	**0.167**	0.684	8.242
Yang et al (2015a)	**0.791**	**2.400**	**0.788**	**0.322**	0.710	4.461
Uricar and Franc (2015)	0.657	7.622	0.677	4.131	0.574	7.957
Xiao et al (2015)	0.760	5.899	**0.782**	**3.845**	0.695	7.379
Rajamanoharan and Cootes (2015)	0.735	6.557	0.717	3.906	0.659	8.289
Wu and Ji (2015)	0.674	13.925	0.732	5.601	0.602	13.161

Colouring denotes the methods' performance ranking per category: 🟥 first 🟩 second 🟦 third 🟧 fourth 🟪 fifth

The area under the curve (AUC) and failure rate are reported. The top five performing curves are highlighted for each video category

Sect. 4.2.2) differs from that of the original competition. This was intended to improve the robustness of the results with respect to variation in pose. Also, as noted in Sect. 4.2, the 300 VW annotations have been corrected and thus this experiment represents updated results for the 300 VW competitors. The results indicate that Yang et al. (2015a) outperforms the rest of the methods for the videos of categories 1 and 2, whereas the deep network of Zhang et al. (2016) combined with CFSS and Kalman smoothing or initialisation from previous are the top performing for the challenging videos of category 3. Moreover, it becomes evident that methodologies which employ face detection dominate category 3, whereas in categories 1 and 2 the model free trackers dominate.

5 Discussion and Conclusions

In Sect. 4 we presented a number of experiments on deformable tracking of sequences containing a single face. We investigated the performance of state-of-the-art face detectors and model free trackers on the recently released 300 VW dataset (see footnote 1). We also devised a number of hybrid systems that attempt to improve the performance of both detectors and trackers with respect to tracking failures. A summary of the proposed experiments are given in Table 4.

Overall, it appears that modern detectors are capable of handling videos of the complexity provided by the

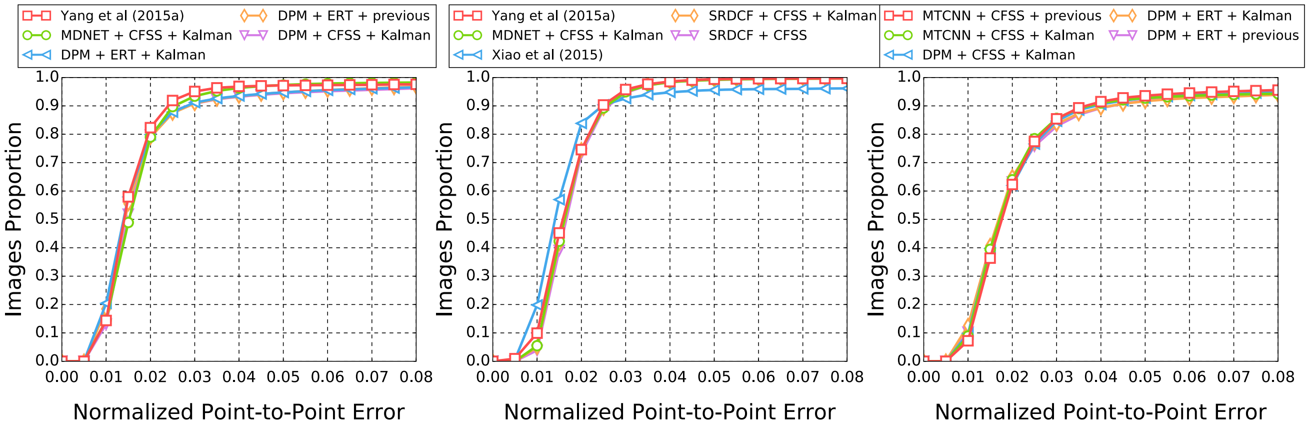


Fig. 13 Comparison between the best methods of Sects. 4.3–4.7 and the participants of the 300 VW challenge by Shen et al. (2015). The top five methods are shown and are coloured *red, blue, green, orange and purple*, respectively. Please see Table 11 for a full summary (Color figure online)

300 VW dataset. This supports the most commonly proposed deformable face tracking methodology that couples a detector with a landmark localisation algorithm. More interestingly, it appears that modern model free trackers are also highly capable of tracking videos that contain variations in pose, expression and illumination. This is particularly evident in the videos of category 2 where the model free trackers perform the best. The performance on the videos of category 2 is likely due to the decreased amount of pose variation in comparison to the other two categories. Category 2 contains many illumination variations which model free trackers appear invariant to. Our work also supports the most recent model free tracking benchmarks (Kristan et al. 2015 and Wu et al. 2015) which have demonstrated that DCF-based trackers are currently the most competitive along with the deep neural network approaches. However, the performance of the trackers does deteriorate significantly in category 3 which supports the categorisation of these videos in the 300 VW as the most difficult category. The difficulty in the videos of category 3 largely stems from the amount of pose variation present, which both detectors and model free trackers struggle with.

The DPM detector provided by Mathias et al. (2014) is very robust across a variety of poses and illumination conditions. The more recent face detector of Zhang et al. (2016) outperforms the rest employed methods in the challenging category 3, however it seems less robust than the DPM detector in the easier categories. The recent advances in the model free trackers, dictate the MDNET tracker of Nam and Han (2016) as a top performing method, which outperforms the pre-trained detectors in the first two categories. MDNET belongs to the discriminatively learned Convolutional Neural Networks trackers with their architecture having several shared CNN layers along with a branched last layer during the training. During the inference, the last layer is discarded and a new layer that is updated online is added. This online update capability of the last layer makes the tracker very robust to abrupt changes and a top performing method in all tracking benchmarks. The SRDCF tracker of Danelljan et al. (2015) from the category of trackers with discriminatively learned correlation filters (DCF) consists an alternative top performing method. DCF trackers are currently a very popular method of choice for bounding box based tracking. They capitalise on a periodic assumption of the training samples to efficiently learn a classifier on all patches in the target neighborhood. Nevertheless, the periodic assumption may introduce unwanted boundary effects, which severely degrade the quality of the tracking model. SRDCF incorporates a spatial regularization component in the learning to penalize correlation filter coefficients depending on their spatial location. The CFSS landmark localisation method of Zhu et al. (2015) outperforms all other considered landmark localisation methods, although the random forest based ERT method of Kazemi and Sullivan (2014) also performed very well. In contrast to the conventional Cascade Regression approaches that iteratively refine an initial shape in a cascaded manner, CFSS explores a diverse shape space and employs a probabilistic heuristic to constrain the finer search in the subsequent cascade levels. The authors argue that this procedure prevents the final solution from being trapped in a local optimum like similar regression techniques. The experimental results support the claim of the authors of Zhu et al. (2015) as the videos contain very challenging pose variations.

The stable performance of both the best model free trackers and detectors on these videos is further demonstrated by the minimal improvement gained from the proposed hybrid systems. Neither reinitialisation from the previous frame (Sect. 4.4), nor the failure detection methodology proposed (Sect. 4.6) improved the best performing methods with any significance. Such hybrid systems could be very useful, though, in case of person re-appearance, multiple person cross-overs. Furthermore, smoothing the facial shapes across the sequences (Kalman) also had a very minimal positive improvement, which can be attributed to the human factor, nonetheless the usage of this smoothing could be more useful for reducing the amount of jiterring in consecutive frames.

In comparison to the recent results of the 300 VW competition (Shen et al. 2015), our review of combinations of modern state-of-the-art detectors and trackers found that very strong performance can be obtained through fairly simple deformable tracking schemes. In fact, only the work of Yang et al. (2015a) outperforms our best performing methods in the easier categories of 1 and 2, while the difference shown by Fig. 13 appears to be marginal. However, the overall results show that, particularly for videos that contain significant pose, there are still improvements to be made.

To summarise, there are a number of important issues that must be tackled in order to improve deformable face tracking:

1. Pose is still a challenging issue for landmark localisation methods. In fact, the videos of 300 VW do not even exhibit the full range of possible facial pose as they do not contain profile faces. The challenges of considering profile faces have yet to be adequately addressed and have not be verified with respect to current state-of-the-art benchmarks.
2. In this work, we only consider videos that contain a single visible face. However, there are many scenarios in which multiple faces may be present and this represents further challenges to deformable tracking. Detectors for example, are particularly vulnerable to multi-object tracking scenarios as they require extending with the ability to determine whether the object being localised is the same as in the previous frame.
3. It is very common for objects to leave the frame of the camera during a sequence, and then reappear. Few model free trackers are robust to reinitialisation after an object has disappeared and then reappeared. When combined with multiple objects, this scenario becomes particularly challenging as it requires a re-identification step in order to verify whether the object to be tracked is one that was seen before.

We believe that deformable face tracking is a very exciting line of research and future advances on the field can have an important impact on several areas of Computer Vision.

Acknowledgements GC was supported by EPSRC DTA award at Imperial College London, as well as from the EPSRC project ADAMANT (EP/L026813/1). The work of PS and EA was funded by the European Community Horizon 2020 [H2020/2014-2020] under Grant Agreement No. 688520 (TeSLA). The work of S. Zafeiriou was funded by the FiDiPro program of Tekes (Project No. 1849/31/2015), as well as from EPSRC Programme Grant FACER2VM (EP/N007743/1).

References

Adam, A., Rivlin, E., & Shimshoni, I. (2006). Robust fragments-based tracking using the integral histogram. In *IEEE proceedings of international conference on computer vision and pattern recognition (CVPR)* (Vol. 1, pp. 798–805). IEEE.

Alabort-i-Medina, & J., Zafeiriou, S. (2014). Bayesian active appearance models. In *IEEE proceedings of international conference on computer vision and pattern recognition (CVPR)* (pp. 3438–3445).

Alabort-i-Medina, J., & Zafeiriou, S. (2015). Unifying holistic and parts-based deformable model fitting. In *IEEE proceedings of international conference on computer vision and pattern recognition (CVPR)* (pp. 3679–3688).

Alabort-i-Medina, J., Antonakos, E., Booth, J., Snape, P., & Zafeiriou, S. (2014). Menpo: A comprehensive platform for parametric image alignment and visual deformable models. In *Proceedings of ACM international conference on multimedia (ACM'MM)* (pp. 679–682). ACM (Code http://www.menpo.org/, Status: Online; accessed June 2, 2016).

Allen, J. G., Xu, R. Y., & Jin, J. S. (2004). Object tracking using camshift algorithm and multiple quantized feature spaces. In *Proceedings of the Pan-Sydney area workshop on visual information processing* (pp. 3–7). Australian Computer Society, Inc.

Amberg, B. (2011). *Editing faces in videos*. PhD thesis, University of Basel.

Amberg, B., Blake, A., & Vetter, T. (2009). On compositional image alignment, with an application to active appearance models. In *IEEE proceedings of international conference on computer vision and pattern recognition (CVPR)* (pp. 1714–1721). IEEE.

Antonakos, E., Alabort-i-Medina, J., Tzimiropoulos, G., & Zafeiriou, S. (2014). Hog active appearance models. *In IEEE proceedings of international conference on image processing (ICIP)* (pp. 224–228).

Antonakos, E., Alabort-i-Medina, J., & Zafeiriou, S. (2015a). Active pictorial structures. In *IEEE proceedings of international conference on computer vision and pattern recognition (CVPR)* (pp. 5435–5444).

Antonakos, E., Alabort-i-Medina, J., Tzimiropoulos, G., & Zafeiriou, S., (2015b). Feature-based lucas-kanade and active appearance models. *IEEE Transactions in Image Processing (TIP)*, *24*(9), 2617–2632.

Arandjelović, R., & Zisserman, A. (2012). Three things everyone should know to improve object retrieval. In *IEEE proceedings of international conference on computer vision and pattern recognition (CVPR)* (pp. 2911–2918). IEEE.

Asthana, A., Zafeiriou, S., Cheng, S., & Pantic, M. (2014). Incremental face alignment in the wild. In *IEEE proceedings of international conference on computer vision and pattern recognition (CVPR)* (pp. 1859–1866).

Asthana, A., Zafeiriou, S., Tzimiropoulos, G., Cheng, S., & Pantic, M. (2015). From pixels to response maps: Discriminative image filtering for face alignment in the wild. *IEEE Transactions on Pattern Analysis and Machine Intelligence (T-PAMI)*, *37*(6), 1312–1320.

Babenko, B., Yang, M. H., & Belongie, S. (2011). Robust object tracking with online multiple instance learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence (T-PAMI), 33*(8), 1619–1632. doi:10.1109/TPAMI.2010.226

Baker, S., & Matthews, I. (2004). Lucas-kanade 20 years on: A unifying framework. *International Journal of Computer Vision (IJCV)*, *56*(3), 221–255.

Balan, A. O., & Black, M. J. (2006). An adaptive appearance model approach for model-based articulated object tracking. In *IEEE proceedings of international conference on computer vision and pattern recognition (CVPR)* (vol 1, pp. 758–765). IEEE.

Barbu, A., Lay, N., & Gramajo, G. (2014). *Face detection with a 3d model*. arXiv preprint arXiv:1404.3596.

Basu, S., Essa, I., & Pentland, A. (1996). Motion regularization for model-based head tracking. In *IEEE international conference on pattern recognition (ICPR)* (vol 3, pp. 611–616). IEEE.

Bay, H., Ess, A., Tuytelaars, T., & Van Gool, L. (2008). Speeded-up robust features (surf). *Computer Vision and Image Understanding*, *110*(3), 346–359.

Belhumeur, P. N., Jacobs, D. W., Kriegman, D. J., & Kumar, N. (2013). Localizing parts of faces using a consensus of exemplars. *IEEE Transactions on Pattern Analysis and Machine Intelligence (T-PAMI)*, *35*(12), 2930–2940.

Bertinetto, L., Valmadre, J., Golodetz, S., Miksik, O., & Torr, P. H. S. (2016a). Staple: Complementary learners for real-time tracking. In *IEEE proceedings of international conference on computer vision and pattern recognition (CVPR)*, IEEE. (Code: https://github.com/bertinetto/staple, Status: Online; accessed August 18 , 2016).

Bertinetto, L., Valmadre, J., Henriques, J. F., Vedaldi, A., & Torr, P. (2016b). *Fully-convolutional siamese networks for object tracking*. arXiv preprint arXiv:1606.09549.

Best-Rowden, L., Klare, B., Klontz, J., & Jain, A. K. (2013). Video-to-video face matching: Establishing a baseline for unconstrained face recognition. In *IEEE sixth international conference on biometrics: Theory, applications and systems (BTAS)* (pp. 1–8). IEEE.

Black, M. J., & Jepson, A. D. (1998). Eigentracking: Robust matching and tracking of articulated objects using a view-based representation. *International Journal of Computer Vision (IJCV)*, *26*(1), 63–84.

Black, M. J., & Yacoob, Y. (1995). Tracking and recognizing rigid and non-rigid facial motions using local parametric models of image motion. In *IEEE Proceedings of International Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 374–381).

Bozdaği, G., Tekalp, A. M., & Onural, L. (1994). 3-d motion estimation and wireframe adaptation including photometric effects for model-based coding of facial image sequences. *IEEE Transactions on Circuits and Systems for Video Technology*, *4*(3), 246–256.

Bradski, G. (2000). *The opencv library*. Dr Dobb's Journal of Software Tools (Code: http://opencv.org, Status: Online; accessed June 2, 2016).

Bradski, G. R. (1998a). Computer vision face tracking as a component of a perceptual user interface. In *Proceedings IEEE workshop on applications of computer vision*, Princeton, NJ, October 1998 (pp. 214–219).

Bradski, G. R., & (1998b). Real time face and object tracking as a component of a perceptual user interface. In *4th IEEE workshop on applications of computer vision, WACV'98* (pp. 214–219). IEEE.

Burgos-Artizzu, X. P., Perona, P., & Dollár, P. (2013). Robust face landmark estimation under occlusion. In *IEEE proceedings of international conference on computer vision (ICCV)*.

Cai, Q., Gallup, D., Zhang, C., & Zhang, Z. (2010). 3d deformable face tracking with a commodity depth camera. In *Proceedings of European conference on computer vision (ECCV)* (pp. 229–242). Springer.

Campbell, K. L. (2016). *Transportation Research Board of the National Academies of Science. The 2nd strategic highway research program naturalistic driving study dataset.*https://insight.shrp2nds.us/, (Online; accessed June 2, 2016).

Cao, X., Wei, Y., Wen, F., & Sun, J. (2014). Face alignment by explicit shape regression. *International Journal of Computer Vision (IJCV)*, *107*(2), 177–190.

Chen, D., Ren, S., Wei, Y., Cao, X., & Sun, J.(2014). Joint cascade face detection and alignment. In *Proceedings of European conference on computer vision (ECCV)* (pp. 109–122). Springer.

Chrysos, G., Antonakos, E., Zafeiriou, S., & Snape, P. (2015). Offline deformable face tracking in arbitrary videos. In *IEEE proceedings of international conference on computer vision, 300 videos in the wild (300-VW): Facial landmark tracking in-the-wild challenge & workshop (ICCV-W)*.

Colmenarez, A., Frey, B., & Huang, T. S. (1999). Detection and tracking of faces and facial features. In *IEEE proceedings of international conference on image processing (ICIP)* (vol 1, pp. 657–661). IEEE.

Comaniciu, D., & Meer, P. (1999). Mean shift analysis and applications. In *IEEE proceedings of international conference on computer vision (ICCV)* (vol 2, pp 1197–1203). IEEE.

Comaniciu, D., Ramesh, V., & Meer, P. (2000). Real-time tracking of non-rigid objects using mean shift. In *IEEE proceedings of international conference on computer vision and pattern recognition (CVPR)* (vol 2, pp 142–149). IEEE.

Cootes, T. F. (2016). Talking face video. http://personalpages.manchester.ac.uk/staff/timothy.f.cootes/data/talking_face/talking_face.html, (Online; accessed June 2, 2016).

Cootes, T. F., Taylor, C. J., Cooper, D. H., & Graham, J. (1995). Active shape models-their training and application. *Computer vision and image understanding*, *61*(1), 38–59.

Cootes, T. F., Edwards, G. J., & Taylor, C. J. (2001). Active appearance models. *IEEE Transactions on Pattern Analysis and Machine Intelligence (T-PAMI)*, *23*(6), 681–685.

Crowley, J. L., & Berard, F. (1997). Multi-modal tracking of faces for video communications. In *IEEE Proceedings of International Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 640–645). IEEE.

Dalal, N., & Triggs, B. (2005). Histograms of oriented gradients for human detection. In *IEEE proceedings of international conference on computer vision and pattern recognition (CVPR)* (pp. 886–893).

Danelljan, M., Häger, G., Khan, F. S., & Felsberg, M. (2014). Accurate scale estimation for robust visual tracking. In *Proceedings of british machine vision conference (BMVC)*.

Danelljan, M., Häger, G., Shahbaz Khan, F., & Felsberg, M. (2015). Learning spatially regularized correlation filters for visual tracking. In *IEEE Proceedings of International Conference on Computer Vision (ICCV)* (pp. 4310–4318). (Code: https://www.cvl.isy.liu.se/en/research/objrec/visualtracking/regvistrack/, Status: Online; accessed June 2, 2016).

Danelljan, M., Robinson, A., Khan, F. S., & Felsberg, M. (2016). Beyond correlation filters: Learning continuous convolution operators for visual tracking. In *Proceedings of European Conference on Computer Vision (ECCV)* (pp. 472–488). (Code: https://github.com/martin-danelljan/Continuous-ConvOp, Status: Online; accessed December 22, 2016).

Decarlo, D., & Metaxas, D. (2000). Optical flow constraints on deformable models with applications to face tracking. *International Journal of Computer Vision (IJCV)*, *38*(2), 99–127.

Dedeoğlu, G., Kanade, T., & Baker, S. (2007). The asymmetry of image registration and its application to face tracking. *IEEE Transactions on Pattern Analysis and Machine Intelligence (T-PAMI)*, *29*(5), 807–823.

De la Torre, F. (2012). A least-squares framework for component analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence (T-PAMI)*, *34*(6), 1041–1055.

Del Moral, P. (1996). Non-linear filtering: Interacting particle resolution. *Markov processes and related fields*, *2*(4), 555–581.

Dollár, P., Tu, Z., Perona, P., & Belongie, S. (2009). Integral channel features. In *Proceedings of British machine vision conference (BMVC)*.

Dollár, P., Welinder, P., & Perona, P. (2010). Cascaded pose regression. In *IEEE Conference on computer vision and pattern recognition (CVPR)* (pp. 1078–1085). IEEE.

Dornaika, F., & Ahlberg, J. (2004). Fast and reliable active appearance model search for 3-d face tracking. *IEEE Transactions On Systems, Man, and Cybernetics, Part B: Cybernetics*, *34*(4), 1838–1853.

Dubout, C., & Fleuret, F. (2012). Exact acceleration of linear object detectors. In *Proceedings of european conference on computer vision (ECCV)* (pp. 301–311) Springer.

Dubout, C., & Fleuret, F. (2013). Deformable part models with individual part scaling. In *Proceedings of British machine vision conference (BMVC), EPFL-CONF-192393*.

Essa, I., Basu, S., Darrell, T., & Pentland, A. (1996). Modeling, tracking and interactive animation of faces and heads using input from video. In *Proceedings of computer animation* (pp. 68–79).

Essa, I., Pentland, A. P., et al. (1997). Coding, analysis, interpretation, and recognition of facial expressions. *IEEE Transactions on Pattern Analysis and Machine Intelligence (T-PAMI)*, *19*(7), 757–763.

Essa, I. A., & Pentland, A. (1994). A vision system for observing and extracting facial action parameters. In *IEEE proceedings of international conference on computer vision and pattern recognition (CVPR)* (pp. 76–83). IEEE.

Essa, I. A., Darrell, T., & Pentland, A. (1994). Tracking facial motion. In *IEEE proceedings of workshop on motion of non-rigid and articulated objects* (pp. 36–42). IEEE.

Felzenszwalb, P. F., & Huttenlocher, D. P. (2005). Pictorial structures for object recognition. *International Journal of Computer Vision (IJCV)*, *61*(1), 55–79.

Felzenszwalb, P. F., Girshick, R. B., McAllester, D., & Ramanan, D. (2010). Object detection with discriminatively trained part-based models. *IEEE Transactions on Pattern Analysis and Machine Intelligence (T-PAMI)*, *32*(9), 1627–1645.

Fischler, M. A., & Elschlager, R. A. (1973). The representation and matching of pictorial structures. *IEEE Transactions on Computers*, *22*(1), 67–92.

Gao, J., Ling, H., Hu, W., & Xing, J. (2014). Transfer learning based visual tracking with gaussian processes regression. In *Proceedings of European Conference on Computer Vision (ECCV)* (pp. 188–203). Springer. (Code: http://www.dabi.temple.edu/~hbling/code/TGPR.htm, Status: Online; accessed December 4, 2016).

Ghiasi, G., & Fowlkes, C. (2014). Occlusion coherence: Localizing occluded faces with a hierarchical deformable part model. In *IEEE proceedings of international conference on computer vision and pattern recognition (CVPR)* (pp. 2385–2392).

Göktürk, S. B., & Tomasi, C. (2004). 3d head tracking based on recognition and interpolation using a time-of-flight depth sensor. In *IEEE proceedings of international conference on computer vision and pattern recognition (CVPR)* (vol 2, pp 2–211). IEEE.

Gordon, N. J., Salmond, D. J., & Smith, A. F. (1993). Novel approach to nonlinear/non-gaussian bayesian state estimation. *Radar and Signal Processing, IEE Proceedings F, IET*, *140*, 107–113.

Grabner, H., Grabner, M., & Bischof, H. (2006). Real-time tracking via on-line boosting. In *Proceedings of British machine vision conference (BMVC)* (vol 5, p. 6).

Gross, R., Matthews, I., Cohn, J., Kanade, T., & Baker, S. (2010). Multi-pie. *Image and Vision Computing*, *28*(5), 807–813.

Hare, S., Saffari, A., & Torr, P. H. (2011). Struck: Structured output tracking with kernels. In *IEEE proceedings of international conference on computer vision (ICCV)* (pp 263–270). IEEE. (Code: http://www.samhare.net/research/struck, Status: Online; accessed June 2, 2016).

Hare, S., Saffari, A., & Torr, P. H. (2012). Efficient online structured output learning for keypoint-based object tracking. In *IEEE proceedings of international conference on computer vision and pattern recognition (CVPR)* (pp. 1894–1901). IEEE.

Heisele, B., Serre, T., Prentice, S., & Poggio, T. (2003). Hierarchical classification and feature reduction for fast face detection with support vector machines. *Pattern Recognition*, *36*(9), 2007–2017.

Henriques, J. F., Caseiro, R., Martins, P., & Batista, J. (2015). High-speed tracking with kernelized correlation filters. *IEEE Transactions on Pattern Analysis and Machine Intelligence (T-PAMI) 37,*(3): 583–596, (Code: https://github.com/joaofaro/KCFcpp, Status: Online; accessed June 2, 2016).

Hjelmås, E., & Low, B. K. (2001). Face detection: A survey. *Computer Vision and Image Understanding*, *83*(3), 236–274.

Hu, P., & Ramanan, D. (2016). *Finding tiny faces*. arXiv preprint arXiv:1612.04402 (Code: https://www.cs.cmu.edu/~peiyunh/tiny/, Status: Online; accessed December 24, 2016).

Huang, G. B., Ramesh, M., Berg, T., & Learned-Miller, E. (2007). *Labeled faces in the wild: A database for studying face recognition in unconstrained environments*. Technical Report 07-49, University of Massachusetts, Amherst.

Isard, M., & Blake, A. (1996). Contour tracking by stochastic propagation of conditional density. In *Proceedings of European conference on computer vision (ECCV)* (pp. 343–356). (Code: https://github.com/gnebehay/SIR-PF, Status: Online; accessed December 23, 2016).

Isard, M., & Blake, A. (1998). Condensationconditional density propagation for visual tracking. *International Journal of Computer Vision (IJCV)*, *29*(1), 5–28.

Jain, V., & Learned-Miller, E. (2010). *Fddb: A benchmark for face detection in unconstrained settings*. Technical Report UM-CS-2010-009, University of Massachusetts, Amherst.

Jepson, A. D., Fleet, D. J., & El-Maraghi, T. F. (2003). Robust online appearance models for visual tracking. *IEEE Transactions on Pattern Analysis and Machine Intelligence (T-PAMI)*, *25*(10), 1296–1311.

Jun, B., Choi, I., & Kim, D. (2013). Local transform features and hybridization for accurate face and human detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence (T-PAMI)*, *35*(6), 1423–1436.

Jurie, F. (1999). A new log-polar mapping for space variant imaging.: Application to face detection and tracking. *Pattern Recognition*, *32*(5), 865–875.

Kalal, Z., Mikolajczyk, K., & Matas, J. (2010a) Face-tld: Tracking-learning-detection applied to faces. In *IEEE proceedings of international conference on image processing (ICIP)* (pp 3789–3792).

Kalal, Z., Mikolajczyk, K., & Matas, J. (2010b). Forward-backward error: Automatic detection of tracking failures. In *IEEE international conference on pattern recognition (ICPR)* (pp 2756–2759). IEEE.

Kalal, Z., Mikolajczyk, K., & Matas, J. (2012). Tracking-learning-detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence (T-PAMI), 34*(7):1409–1422, (Code: https://github.com/zk00006/OpenTLD, Status: Online; accessed June 2, 2016).

Kalman, R. E. (1960). A new approach to linear filtering and prediction problems. *Journal of Basic Engineering*, *82*(1), 35–45.

Kazemi, V., & Sullivan, J. (2014). One millisecond face alignment with an ensemble of regression trees. In *IEEE proceedings of international conference on computer vision and pattern recognition (CVPR)* (pp. 1867–1874).

Kim, M., Kumar, S., Pavlovic, V., & Rowley, H. (2008). Face tracking and recognition with visual constraints in real-world videos. In *IEEE proceedings of international conference on computer vision and pattern recognition (CVPR)* (pp. 1–8) IEEE.

King, D. E. (2009). Dlib-ml: A machine learning toolkit. *The Journal of Machine Learning Research, 10*:1755–1758, (Code: http://dlib.net/, Status: Online; accessed June 2, 2016).

King, D. E. (2015). Max-margin object detection. arXiv preprint arXiv:1502.00046.

Klare, B. F., Klein, B., Taborsky, E., Blanton, A., Cheney, J., Allen, K., Grother, P., Mah, A., Burge, M., & Jain, A. K. (2015). Pushing the frontiers of unconstrained face detection and recognition: Iarpa janus benchmark a. In *IEEE proceedings of international conference on computer vision and pattern recognition (CVPR)* (pp 1931–1939) IEEE.

Koelstra, S., Pantic, M., & Patras, I. Y. (2010). A dynamic texture-based approach to recognition of facial actions and their temporal models. *IEEE Transactions on Pattern Analysis and Machine Intelligence (T-PAMI)*, *32*(11), 1940–1954.

Kokiopoulou, E., Chen, J., & Saad, Y. (2011). Trace optimization and eigenproblems in dimension reduction methods. *Numerical Linear Algebra with Applications*, *18*(3), 565–602.

Köstinger, M., Wohlhart, P., Roth, P. M., & Bischof, H. (2011). Annotated facial landmarks in the wild: A large-scale, real-world database for facial landmark localization. In *IEEE proceedings of international conference on computer vision workshops (ICCV'W)* (pp. 2144–2151).

Köstinger, M., Wohlhart, P., Roth, P. M., & Bischof, H. (2012). Robust face detection by simple means. In *DAGM 2012 CVAW workshop*.

Koukis, V., Venetsanopoulos, C., & Koziris, N. (2013). Okeanos: Building a cloud, cluster by cluster. *IEEE Internet Computing*, *17*(3), 67–71.

Kristan, M., Pflugfelder, R., Leonardis, A., Matas, J., Porikli, F., Čehovin, L., Nebehay, G., et al. (2013). The visual object tracking vot2013 challenge results. In *IEEE Proceedings of international conference on computer vision workshops (ICCV'W)*.

Kristan, M., Pflugfelder, R., Leonardis, A., Matas, J., Čehovin, L., Nebehay, G., et al. (2014). The visual object tracking vot2014 challenge results. In *Proceedings of European conference on computer vision workshops (ECCV'W)*, http://www.votchallenge.net/vot2014/program.html.

Kristan, M., Matas, J., Leonardis, A., Felsberg, M., Čehovin, L., Fernandez, G., et al. (2015). The visual object tracking vot2015 challenge results. In *IEEE proceedings of international conference on computer vision workshops (ICCV'W)*.

Kristan, M., Matas, J., Leonardis, A., Vojíř, T., Pflugfelder, R., Fernandez, G., et al. (2016). A novel performance evaluation methodology for single-target trackers. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 38*(11), 2137–2155.

Kumar, V., Namboodiri, A., & Jawahar, C. (2015). Visual phrases for exemplar face detection. In *IEEE proceedings of international conference on computer vision (ICCV)* (pp. 1994–2002). (Code: http://cvit.iiit.ac.in/projects/exemplar/, Status: Online; accessed December 24, 2016).

La Cascia, M., Sclaroff, S., & Athitsos, V. (2000). Fast, reliable head tracking under varying illumination: An approach based on registration of texture-mapped 3d models. *IEEE Transactions on Pattern Analysis and Machine Intelligence (T-PAMI)*, *22*(4), 322–336.

Lanitis, A., Taylor, C. J., & Cootes, T. F. (1995). A unified approach to coding and interpreting face images. In *IEEE proceedings of international conference on computer vision and pattern recognition (CVPR)* (pp. 368–373).

Le, V., Brandt, J., Lin, Z., Bourdev, L., & Huang, T. S. (2012). Interactive facial feature localization. In *Proceedings of European conference on computer vision (ECCV)* (pp. 679–692). Springer.

Learned-Miller, E., Huang, G. B., RoyChowdhury, A., Li, H., & Hua, G. (2016). Labeled faces in the wild: A survey. In *Advances in face detection and facial image analysis* (pp. 189–248). Springer International Publishing.

Levey, A., & Lindenbaum, M. (2000). Sequential karhunen-loeve basis extraction and its application to images. *IEEE Transactions on Image Processing*, *9*(8), 1371–1374.

Li, A., Lin, M., Wu, Y., Yang, M. H., & Yan, S. (2016a). Nus-pro: A new visual tracking challenge. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 38*(2), 335–349.

Li, A., Lin, M., Wu, Y., Yang, M. H., & Yan, S. (2016b). *Nus-pro tracking challenge*. http://www.lv-nus.org/pro/nus_pro.html, (Online; accessed June 2, 2016).

Li, H., Roivainen, P., & Forchheimer, R. (1993). 3-d motion estimation in model-based facial image coding. *IEEE Transactions on Pattern Analysis and Machine Intelligence (T-PAMI)*, *15*(6), 545–555.

Li, H., Hua, G., Lin, Z., Brandt, J., & Yang, J. (2013a). Probabilistic elastic part model for unsupervised face detector adaptation. In *IEEE proceedings of international conference on computer vision (ICCV)* (pp. 793–800).

Li, H., Lin, Z., Brandt, J., Shen, X., & Hua, G. (2014). Efficient boosted exemplar-based face detection. In *IEEE proceedings of international conference on computer vision and pattern recognition (CVPR)* (pp. 1843–1850).

Li, H., Lin, Z., Shen, X., Brandt, J., & Hua, G. (2015a). A convolutional neural network cascade for face detection. In *IEEE Proceedings of international conference on computer vision and pattern recognition (CVPR)* (pp. 5325–5334).

Li, J., & Zhang, Y. (2013). Learning surf cascade for fast and accurate object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 3468–3475).

Li, J., Wang, T., & Zhang, Y. (2011). Face detection using surf cascade. In *IEEE proceedings of international conference on computer vision workshops (ICCV'W)* (pp. 2183–2190). IEEE.

Li, S. Z., Zhu, L., Zhang, Z., Blake, A., Zhang, H., & Shum, H. (2002). Statistical learning of multi-view face detection. In *Proceedings of European conference on computer vision (ECCV)* (pp. 67–81). Springer.

Li, X., Hu, W., Shen, C., Zhang, Z., Dick, A., & Hengel, A. V. D. (2013b). A survey of appearance models in visual object tracking. *ACM Transactions on Intelligent Systems and Technology (TIST)*, *4*(4), 58.

Li, Y., Gong, S., & Liddell, H. (2000). Support vector regression and classification based multi-view face detection and recognition. In *IEEE proceedings of international conference on automatic face and gesture recognition (FG)* (pp. 300–305) IEEE.

Li, Y., Zhu, J., & Hoi, S. C. (2015b). Reliable patch trackers: Robust visual tracking by exploiting reliable patches. In *IEEE Proceedings of International Conference on Computer Vision and Pattern Recognition (CVPR)* (pp 353–361). (Code: https://github.com/ihpdep/rpt, Status: Online; accessed June 2, 2016).

Liao, S., Jain, A. K., & Li, S. Z. (2016). A fast and accurate unconstrained face detector. *IEEE Transactions on Pattern Analysis and Machine Intelligence (T-PAMI), 38*(2):211–223, (Code: http://www.cbsr.ia.ac.cn/users/scliao/projects/npdface/, Status: Online; accessed December 24, 2016).

Liwicki, S., Zafeiriou, S., & Pantic, M. (2012a). Incremental slow feature analysis with indefinite kernel for online temporal video segmentation. In *Asian conference on computer vision (ACCV)* (pp 162–176). Springer.

Liwicki, S., Zafeiriou, S., Tzimiropoulos, G., & Pantic, M. (2012b). Efficient online subspace learning with an indefinite kernel for visual tracking and recognition. *IEEE Transactions on Neural Networks and Learning Systems (T-NN), 23*(10):1624–1636.

Liwicki, S., Tzimiropoulos, G., Zafeiriou, S., & Pantic, M. (2013). Euler principal component analysis. *International Journal of Computer Vision (IJCV)*, *101*(3), 498–518.

Liwicki, S., Zafeiriou, S. P., & Pantic, M. (2015). Online kernel slow feature analysis for temporal video segmentation and tracking. *IEEE Transactions in Image Processing (TIP)*, *24*(10), 2955–2970.

Liwicki, S., Zafeiriou, S., Tzimiropoulos, G., & Pantic, M. (2016). *Annotated face videos*. http://www.robots.ox.ac.uk/~stephan/dikt/, (Online; accessed June 2, 2016).

Lowe, D. G. (1999). Object recognition from local scale-invariant features. In *IEEE proceedings of international conference on computer vision (ICCV)* (pp. 1150–1157).

Ma, C., Yang, X., Zhang, C., & Yang, M. H. (2015). Long-term correlation tracking. In *IEEE proceedings of international conference on computer vision and pattern recognition (CVPR)* (pp. 5388–5396). IEEE (Code: https://github.com/chaoma99/lct-tracker, Status: Online; accessed August 18, 2016).

Malciu, M., & Prěteux, F. (2000). A robust model-based approach for 3d head tracking in video sequences. In *IEEE proceedings of international conference on automatic face and gesture recognition (FG)* (pp 169–174). IEEE.

Mathias, M., Benenson, R., Pedersoli, M., & Van Gool, L. (2014). Face detection without bells and whistles. In *Proceedings of European conference on computer vision (ECCV)* (pp 720–735) Springer.

Matthews, I., & Baker, S. (2004). Active appearance models revisited. *International Journal of Computer Vision (IJCV)*, *60*(2), 135–164.

Matthews, I., Ishikawa, T., & Baker, S. (2004). The template update problem. *IEEE Transactions on Pattern Analysis and Machine Intelligence (T-PAMI)*, *26*(6), 810–815.

Mei, X., & Ling, H. (2011). Robust visual tracking and vehicle classification via sparse representation. *IEEE Transactions on Pattern Analysis and Machine Intelligence (T-PAMI)*, *33*(11), 2259–2272.

Mita, T., Kaneko, T., & Hori, O. (2005). Joint haar-like features for face detection. *IEEE Proceedings of International Conference on Computer Vision (ICCV)*, *2*, 1619–1626.

Nam, H., & Han, B. (2016). Learning multi-domain convolutional neural networks for visual tracking. In *IEEE proceedings of international conference on computer vision and pattern recognition (CVPR)*, IEEE, (Code: https://github.com/HyeonseobNam/MDNet, Status: Online; accessed August 18, 2016).

Nebehay, G., & Pflugfelder, R. (2015). Clustering of static-adaptive correspondences for deformable object tracking. In *IEEE proceedings of international conference on computer vision and pattern recognition (CVPR)*, IEEE (Code: https://github.com/gnebehay/CppMT, Status: Online; accessed June 2, 2016).

Ning, J., Yang, J., Jiang, S., Zhang, L., & Yang, M. H. (2016). Object tracking via dual linear structured svm and explicit feature map. In *IEEE proceedings of international conference on computer vision and pattern recognition (CVPR)*, (Code: www4.comp.polyu.edu.hk/~cslzhang/code/DLSSVM_CVPR.zip, Status: Online; accessed August 18, 2016)

Ojala, T., Pietikäinen, M., & Mäenpää, T. (2002). Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. *IEEE Transactions on Pattern Analysis and Machine Intelligence (T-PAMI)*, *24*(7), 971–987.

Oliver, N., Pentland, A. P., & Berard, F. (1997). Lafter: Lips and face real time tracker. In *IEEE proceedings of international conference on computer vision and pattern recognition (CVPR)* (pp. 123–129)

Orozco, J., Rudovic, O., Gonzàlez, J., & Pantic, M. (2013). Hierarchical on-line appearance-based tracking for 3d head pose, eyebrows, lips, eyelids and irises. *Image and Vision Computing*, *31*(4), 322–340.

Osadchy, M., Cun, Y. L., & Miller, M. L. (2007). Synergistic face detection and pose estimation with energy-based models. *Journal of Machine Learning Research*, *8*, 1197–1215.

Papandreou, G., & Maragos, P. (2008). Adaptive and constrained algorithms for inverse compositional active appearance model fitting. In *IEEE proceedings of international conference on computer vision and pattern recognition (CVPR)*, IEEE, (pp. 1–8).

Parkhi, O. M., Vedaldi, A., & Zisserman, A. (2015). Deep face recognition. *Proceedings of the British Machine Vision*, *1*(3), 6.

Patras, I., & Pantic, M. (2004). Particle filtering with factorized likelihoods for tracking facial features. In *IEEE proceedings of international conference on automatic face and gesture recognition (FG)* (pp. 97–102).

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research, 12*, 2825–2830, (Code: http://scikit-learn.org/, Status: Online; accessed December 22, 2016).

Peng, Y., Ganesh, A., Wright, J., Xu, W., & Ma, Y. (2012). Rasl: Robust alignment by sparse and low-rank decomposition for linearly correlated images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *34*(11), 2233–2246.

Pérez, F., & Granger, B. E. (2007). IPython: a system for interactive scientific computing. *Computing in Science and Engineering 9,* 21–29, (Code: https://ipython.org/, Status: Online; accessed December 22, 2016).

Pernici, F., & Del Bimbo, A. (2014). Object tracking by oversampling local features. *IEEE Transactions on Pattern Analysis and Machine Intelligence (T-PAMI)*, *36*(12), 2538–2551.

Phillips, P. J., Moon, H., Rizvi, S. A., & Rauss, P. J. (2000). The feret evaluation methodology for face-recognition algorithms. *IEEE Transactions on Pattern Analysis and Machine Intelligence (T-PAMI)*, *22*(10), 1090–1104.

Pighin, F., Szeliski, R., & Salesin, D. H. (1999). Resynthesizing facial animation through 3d model-based tracking. In *IEEE Proceedings of International Conference on Computer Vision (ICCV)* (vol 1, pp. 143–150). IEEE.

Poling, B., Lerman, G., & Szlam, A. (2014). Better feature tracking through subspace constraints. In *IEEE proceedings of international conference on computer vision and pattern recognition (CVPR)* (pp. 3454–3461).

Qi, Y., Zhang, S., Qin, L., Yao, H., Huang, Q., & Yang, J. L. M. H. (2016). Hedged deep tracking. In *IEEE proceedings of international conference on computer vision and pattern recognition (CVPR)*. IEEE (Code: https://sites.google.com/site/yuankiqi/hdt/, Status: Online; accessed December 4, 2016).

Qian, R. J., Sezan, M. I., & Matthews, K. E. (1998). A robust real-time face tracking algorithm. In *IEEE proceedings of international conference on image processing (ICIP)* (vol 1, pp 131–135). IEEE.

Rajamanoharan, G., & Cootes, T. (2015). Multi-view constrained local models for large head angle face tracking. In *IEEE proceedings of international conference on computer vision, 300 videos in the wild (300-VW): Facial landmark tracking in-the-wild challenge & workshop (ICCV-W)*.

Ranjan, R., Patel, V. M., & Chellappa, R. (2015). A deep pyramid deformable part model for face detection. *IEEE International Conference on Biometrics Theory* (pp. 1–8). IEEE: Applications and Systems (BTAS).

Rätsch, M., Romdhani, S., & Vetter, T. (2004). Efficient face detection by a cascaded support vector machine using haar-like features. In *Pattern recognition* (pp. 62–70). Springer.

Ren, S., Cao, X., Wei, Y., & Sun, J. (2014). Face alignment at 3000 fps via regressing local binary features. In *IEEE proceedings of international conference on computer vision and pattern recognition (CVPR)* (pp. 1685–1692). IEEE.

Romdhani, S., Torr, P., Schölkopf, B., & Blake, A. (2001). Computationally efficient face detection. In *IEEE proceedings of international conference on computer vision (ICCV)* (vol 2, pp 695–700). IEEE.

Ross, D., Lim, J., Lin, R. S., & Yang, M. H. (2015). *Dudek face sequence*. http://www.cs.toronto.edu/~dross/ivt/ (Online; accessed June 2, 2016).

Ross, D. A., Lim, J., Lin, R. S., & Yang, M. H. (2008). Incremental learning for robust visual tracking. *International Journal of Computer Vision (IJCV) 77,*(1–3), 125–141, (Code: http://www.cs.toronto.edu/~dross/ivt/, Status: Online; accessed June 2, 2016).

Rueckert, D., Sonoda, L. I., Hayes, C., Hill, D. L., Leach, M. O., & Hawkes, D. J. (1999). Nonrigid registration using free-form deformations: Application to breast mr images. *IEEE Transactions on Medical Imaging*, *18*(8), 712–721.

Sagonas, C., Tzimiropoulos, G., Zafeiriou, S., & Pantic, M. (2013a). 300 faces in-the-wild challenge: The first facial landmark localization challenge. In *IEEE proceedings of international conference on computer vision (ICCV-W), 300 faces in-the-wild challenge (300-W)* (pp. 397–403).

Sagonas, C., Tzimiropoulos, G., Zafeiriou, S., & Pantic, M. (2013b). A semi-automatic methodology for facial landmark annotation. In *IEEE proceedings of international conference on computer vision and pattern recognition (CVPR-W), 5th workshop on analysis and modeling of faces and gestures* (pp. 896–903).

Sagonas, C., Panagakis, Y., Zafeiriou, S., & Pantic, M. (2014). Raps: Robust and efficient automatic construction of person-specific deformable models. In *IEEE proceedings of international conference on computer vision and pattern recognition (CVPR)* (pp. 1789–1796).

Sagonas, C., Antonakos, E., Tzimiropoulos, G., Zafeiriou, S., & Pantic, M. (2015). 300 Faces in-the-wild challenge: Database and results. In *Image and vision computing*.

Sakai, T., Nagao, M., & Kanade, T. (1972). Computer analysis and classification of photographs of human faces. In *Proceedings of First USA-JAPAN Computer Conference* (pp. 55–62).

Salti, S., Cavallaro, A., & Stefano, L. D. (2012). Adaptive appearance modeling for video tracking: Survey and evaluation. *IEEE Transactions in Image Processing (TIP)*, *21*(10), 4334–4348.

Saragih, J. M., Lucey, S., & Cohn, J. F. (2011). Deformable model fitting by regularized landmark mean-shift. *International Journal of Computer Vision*, *91*(2), 200–215.

Schneiderman, H., & Kanade, T. (2004). Object detection using the statistics of parts. *International Journal of Computer Vision (IJCV)*, *56*(3), 151–177.

Schroff, F., Kalenichenko, D., & Philbin, J. (2015). Facenet: A unified embedding for face recognition and clustering. In *IEEE proceedings of international conference on computer vision and pattern recognition (CVPR)* (pp. 815–823).

Schwerdt, K., & Crowley, J. L. (2000). Robust face tracking using color. In *IEEE proceedings of international conference on automatic face and gesture recognition (FG)* (pp. 90–95). IEEE.

Sevilla-Lara, L., & Learned-Miller, E. (2012). Distribution fields for tracking. In *IEEE proceedings of international conference on computer vision and pattern recognition (CVPR)* (pp. 1910–1917). IEEE. (Code: http://people.cs.umass.edu/~lsevilla/trackingDF.html, Status: Online; accessed June 2, 2016).

Shen, J., Zafeiriou, S., Chrysos, G., Kossaifi, J., Tzimiropoulos, G., & Pantic, M. (2015). The first facial landmark tracking in-the-wild challenge: Benchmark and results. In *IEEE proceedings of international conference on computer vision, 300 videos in the wild (300-VW): Facial landmark tracking in-the-wild challenge & workshop (ICCV-W)*.

Shen, X., Lin, Z., Brandt, J., & Wu, Y. (2013). Detecting and aligning faces by image retrieval. In *IEEE proceedings of international conference on computer vision and pattern recognition (CVPR)* (pp. 3460–3467).

Smeulders, A. W., Chu, D. M., Cucchiara, R., Calderara, S., Dehghan, A., & Shah, M. (2014). Visual tracking: An experimental survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence (T-PAMI)*, *36*(7), 1442–1468.

Snape, P., Roussos, A., Panagakis, Y., & Zafeiriou, S. (2015). Face flow. In *IEEE proceedings of international conference on computer vision (ICCV)*.

Sobottka, K., & Pitas, I. (1996). Face localization and facial feature extraction based on shape and color information. In *IEEE proceedings of international conference on image processing (ICIP)* (vol 3, pp 483–486). IEEE.

Stern, H., & Efros, B. (2002). Adaptive color space switching for face tracking in multi-colored lighting environments. In *IEEE proceedings of international conference on automatic face and gesture recognition (FG)* (pp 249–254). IEEE.

Sung, J., & Kim, D. (2009). Adaptive active appearance model with incremental learning. *Pattern Recognition Letters*, *30*(4), 359–367.

Sung, J., Kanade, T., & Kim, D. (2008). Pose robust face tracking by combining active appearance models and cylinder head models. *International Journal of Computer Vision (IJCV)*, *80*(2), 260–274.

Taigman, Y., Yang, M., Ranzato, M., & Wolf, L. (2014). Deepface: Closing the gap to human-level performance in face verification. In *IEEE Proceedings of International Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 1701–1708).

Tao, H., & Huang, T. S. (1998). Connected vibrations: A modal analysis approach for non-rigid motion tracking. In *IEEE proceedings of international conference on computer vision and pattern recognition (CVPR)* (pp 735–740). IEEE.

Toyama, K. (1998). *Look, ma-no hands! hands-free cursor control with real-time 3d face tracking. PUI98.*

Tresadern, P. A., Ionita, M. C., & Cootes, T. F. (2012). Real-time facial feature tracking on a mobile device. *International Journal of Computer Vision (IJCV)*, *96*(3), 280–289.

Tzimiropoulos, G. (2015). Project-out cascaded regression with an application to face alignment. In *IEEE proceedings of international conference on computer vision and pattern recognition (CVPR)*. (pp. 3659–3667).

Tzimiropoulos, G., & Pantic, M. (2013). Optimization problems for fast aam fitting in-the-wild. In *IEEE proceedings of international conference on computer vision (ICCV)* (pp. 593–600). IEEE.

Tzimiropoulos, G., & Pantic, M. (2014). Gauss-newton deformable part models for face alignment in-the-wild. In *IEEE Proceedings of International Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 1851–1858).

Tzimiropoulos, G., Alabort-i-Medina, J., Zafeiriou, S., & Pantic, M. (2012). Generic active appearance models revisited. In *Asian conference on computer vision (ACCV)* (pp. 650–663). Springer.

Tzimiropoulos, G., Alabort-i Medina, J., Zafeiriou, S., & Pantic, M. (2014). Active orientation models for face alignment in-the-wild. *IEEE Transactions on Information Forensics and Security*, *9*(12), 2024–2034.

Uricar, M., & Franc, V. (2015). Real-time facial landmark tracking by tree-based deformable part model based detector. In *IEEE proceedings of international conference on computer vision, 300 videos in the wild (300-VW): Facial landmark tracking in-the-wild challenge & workshop (ICCV-W)*.

Vadakkepat, P., Lim, P., De Silva, L. C., Jing, L., & Ling, L. L. (2008). Multimodal approach to human-face detection and tracking. *IEEE Transactions on Industrial Electronics*, *55*(3), 1385–1393.

Viola, P., & Jones, M. (2001). Rapid object detection using a boosted cascade of simple features. In *IEEE proceedings of international conference on computer vision and pattern recognition (CVPR)* (vol 1, pp I–511). IEEE.

Viola, P., & Jones, M. J. (2004). Robust real-time face detection. *International Journal of Computer Vision (IJCV)*, *57*(2), 137–154.

Wang, N., Gao, X., Tao, D., & Li, X. (2014). *Facial feature point detection: A comprehensive survey*. arXiv preprint arXiv:1410.1037.

Wang, P., Ji, Q. (2004). Multi-view face detection under complex scene based on combined svms. In *IEEE International Conference on Pattern Recognition (ICPR)* (pp. 179–182).

Wang, X., Valstar, M., Martinez, B., Haris Khan, M., & Pridmore, T. (2015). Tric-track: Tracking by regression with incrementally learned cascades. In *IEEE proceedings of international conference on computer vision (ICCV)* (pp. 4337–4345).

Wei, X., Zhu, Z., Yin, L., & Ji, Q. (2004). A real time face tracking and animation system. In *IEEE Proceedings of International Con-*

ference on Computer Vision and Pattern Recognition Workshops (CVPR'W) (pp. 71–71). IEEE.

Weise, T., Bouaziz, S., Li, H., & Pauly, M. (2011). Realtime performance-based facial animation. In *ACM transactions on graphics (TOG)* (vol 30, p. 77). ACM.

Wolf, L., Hassner, T., Maoz, I. (2011). Face recognition in unconstrained videos with matched background similarity. In *IEEE proceedings of international conference on computer vision and pattern recognition (CVPR)* (pp. 529–534).

Wu, B., Ai, H., Huang, C., & Lao, S. (2004). Fast rotation invariant multi-view face detection based on real adaboost. In *IEEE proceedings of international conference on automatic face and gesture recognition (FG)* (pp. 79–84). IEEE.

Wu, Y., & Ji, Q. (2015). Shape augmented regression method for face alignment. In *IEEE proceedings of international conference on computer vision, 300 videos in the wild (300-VW): Facial landmark tracking in-the-wild challenge & workshop (ICCV-W)*.

Wu, Y., Shen, B., & Ling, H. (2012). Online robust image alignment via iterative convex optimization. In *IEEE proceedings of international conference on computer vision and pattern recognition (CVPR)* (pp. 1808–1814). IEEE. (Code: https://sites.google.com/site/trackerbenchmark/benchmarks/v10, Status: Online; accessed June 2, 2016).

Wu, Y., Lim, J., & Yang, M. H. (2013). Online object tracking: A benchmark. In *IEEE proceedings of international conference on computer vision and pattern recognition (CVPR)*.

Wu, Y., Lim, J., & Yang, M. H. (2015). Object tracking benchmark. *IEEE Transactions on Pattern Analysis and Machine Intelligence (T-PAMI)*, *37*(9), 1834–1848.

Xiao, J., Baker, S., Matthews, I., & Kanade, T. (2004). Real-time combined 2d+ 3d active appearance models. In *IEEE proceedings of international conference on computer vision and pattern recognition (CVPR)* (pp. 535–542).

Xiao, S., Yan, S., & Kassim, A. (2015). Facial landmark detection via progressive initialization. In *IEEE Proceedings of international conference on computer vision, 300 videos in the wild (300-VW): facial landmark tracking in-the-wild challenge & workshop (ICCV-W)*.

Xiao, Z., Lu, H., & Wang, D. (2014). L2-RLS-based object tracking. *IEEE Transactions on Circuits and Systems for Video Technology*, *24*(8), 1301–1309.

Xiong, X., & De la Torre, F. (2013). Supervised descent method and its applications to face alignment. In *IEEE proceedings of international conference on computer vision and pattern recognition (CVPR)* (pp. 532–539).

Xiong, X., & De la Torre, F. (2015). Global supervised descent method. In *IEEE proceedings of international conference on computer vision and pattern recognition (CVPR)* (pp. 2664–2673).

Yacoob, Y., & Davis, L. S. (1996). Recognizing human facial expressions from long image sequences using optical flow. *IEEE Transactions on Pattern Analysis and Machine Intelligence (T-PAMI)*, *18*(6), 636–642.

Yan, J., Zhang, X., Lei, Z., Yi, D., & Li, S. Z. (2013). Structural models for face detection. In *IEEE Proceedings of International Conference on Automatic Face and Gesture Recognition (FG)* (pp. 1–6). IEEE.

Yan, J., Lei, Z., Wen, L., & Li, S. (2014). The fastest deformable part model for object detection. In *IEEE Proceedings of International Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 2497–2504).

Yang, B., Yan, J., Lei, Z., & Li, S. Z. (2014a). Aggregate channel features for multi-view face detection. In *IEEE international joint conference on biometrics (IJCB)* (pp. 1–8). IEEE.

Yang, F., Lu, H., & Yang, M. H. (2014b). Robust superpixel tracking. *IEEE Transactions in Image Processing (TIP), 23*(4), 1639–1651, (Code: http://www.umiacs.umd.edu/~fyang/spt.html, Status: Online; accessed August 18, 2016).

Yang, H., Shao, L., Zheng, F., Wang, L., & Song, Z. (2011). Recent advances and trends in visual tracking: A review. *Neurocomputing*, *74*(18), 3823–3831.

Yang, J., Deng, J., Zhang, K., & Liu, Q. (2015a). Facial shape tracking via spatio-temporal cascade shape regression. In *IEEE proceedings of international conference on computer vision, 300 videos in the wild (300-VW): Facial landmark tracking in-the-wild challenge & workshop (ICCV-W)*.

Yang, M. H., Kriegman, D. J., & Ahuja, N. (2002). Detecting faces in images: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence (T-PAMI)*, *24*(1), 34–58.

Yang, S., Luo, P., Loy, C. C., & Tang, X. (2015b). From facial parts responses to face detection: A deep learning approach. In *IEEE Proceedings of International Conference on Computer Vision (ICCV)* (pp. 3676–3684).

Yao, R., Shi, Q., Shen, C., Zhang, Y., & Hengel, A. (2013). Part-based visual tracking with online latent structural learning. In *IEEE proceedings of international conference on computer vision and pattern recognition (CVPR)* (pp. 2363–2370).

Zafeiriou, S., Zhang, C., & Zhang, Z. (2015). A survey on face detection in the wild past, present and future. *Computer Vision and Image Understanding*, *138*, 1–24.

Zhang, C., & Zhang, Z. (2010). *A survey of recent advances in face detection*. Technical report, Microsoft Research.

Zhang, C., & Zhang, Z. (2014). Improving multiview face detection with multi-task deep convolutional neural networks. In *IEEE winter conference on applications of computer vision (WACV)* (pp. 1036–1041). IEEE.

Zhang, J., Ma, S., & Sclaroff, S. (2014a). Meem: robust tracking via multiple experts using entropy minimization. In *Proceedings of European conference on computer vision (ECCV)* (pp. 188–203). (Code: http://cs-people.bu.edu/jmzhang/MEEM/MEEM.html, Status: Online; accessed August 18, 2016).

Zhang, K., Zhang, L., Liu, Q., Zhang, D., & Yang, M. H. (2014b). Fast visual tracking via dense spatio-temporal context learning. In *Proceedings of European conference on computer vision (ECCV)* (pp. 127–141). (Code: http://www4.comp.polyu.edu.hk/~cslzhang/STC/STC.htm, Status: Online; accessed August 18, 2016).

Zhang, K., Zhang, L., & Yang, M. H. (2014c). Fast compressive tracking. *IEEE Transactions on Pattern Analysis and Machine Intelligence (T-PAMI), 36*(10), 2002–2015, (Code: http://www4.comp.polyu.edu.hk/~cslzhang/FCT/FCT.htm, Status: Online; accessed June 2, 2016).

Zhang, K., Zhang, Z., Li, Z., & Qiao, Y. (2016). Joint face detection and alignment using multi-task cascaded convolutional networks. *IEEE Signal Processing Letters, 23*(10), 1499–1503, (Code: https://github.com/kpzhang93/MTCNN_face_detection_alignment, Status: Online; accessed December 24, 2016).

Zhang, L., & van der Maaten, L. (2013). Structure preserving object tracking. In *IEEE Proceedings of International Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 1838–1845). IEEE.

Zhang, L., van der Maaten, L. (2014). Preserving structure in model-free tracking. *IEEE Transactions on Pattern Analysis and Machine Intelligence (T-PAMI), 36*(4), 756–769, (Code: http://visionlab.tudelft.nl/spot, Status: Online; accessed June 2, 2016).

Zhang, T., Ghanem, B., Liu, S., & Ahuja, N. (2012). Robust visual tracking via multi-task sparse learning. In *IEEE proceedings of international conference on computer vision and pattern recognition (CVPR)* (pp. 2042–2049). IEEE.

Zhang, T., Liu, S., Ahuja, N., Yang, M. H., & Ghanem, B. (2014d). Robust visual tracking via consistent low-rank sparse learning. *International Journal of Computer Vision (IJCV), 111*(2), 171–190, (Code: http://nlpr-web.ia.ac.cn/mmc/homepage/tzzhang/Project_Tianzhu/zhang_IJCV14/Robust%20Visual%20Tracking%20Via%20Consistent%20Low-Rank%20Sparse.html, Status: Online; accessed June 2, 2016).

Zhang, W., Wang, Q., & Tang, X. (2008). Real time feature based 3-d deformable face tracking. In *Proceedings of European Conference on Computer Vision (ECCV)* (pp. 720–732). Springer.

Zhu, S., Li, C., Loy, C. C., & Tang, X. (2015). Face alignment by coarse-to-fine shape searching. In *IEEE proceedings of international conference on computer vision and pattern recognition (CVPR)* (pp. 4998–5006) (Code: https://github.com/zhusz/CVPR15-CFSS, Status: Online; accessed December 4, 2016).

Zhu, X., & Ramanan, D. (2012). Face detection, pose estimation, and landmark localization in the wild. In: *IEEE Proceedings of International Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 2879–2886). IEEE. (Code: https://www.ics.uci.edu/~xzhu/face, Status: Online; accessed June 2, 2016).