PD^{2}T: Person-specific Detection, Deformable Tracking

Grigorios G. Chrysos and Stefanos Zafeiriou, *Member, IEEE*

Abstract—Face detection/alignment methods have reached a satisfactory state in static images captured under arbitrary conditions. Such methods typically perform (joint) fitting for each frame and are used in commercial applications; however in the majority of the real-world scenarios the dynamic scenes are of interest. We argue that generic fitting per frame is suboptimal (it discards the informative correlation of sequential frames) and propose to learn person-specific statistics from the video to improve the generic results. To that end, we introduce a meticulously studied pipeline, which we name PD^{2}T, that performs person-specific detection and landmark localisation. We carry out extensive experimentation with a diverse set of i) generic fitting results, ii) different objects (human faces, animal faces) that illustrate the powerful properties of our proposed pipeline and experimentally verify that PD^{2}T outperforms all the compared methods.

Index Terms—Deformable models, adaptive tracking, person-specific learning, long-term tracking

1 Introduction

Significant progress has been made in static facial imagery, e.g. in face detection ([1], [2], [3]), landmark localisation ([4], [5], [6], [7]). Nevertheless, in a wide range of applications such as lip-reading, expression analysis, surveillance, commercial cameras' tracking the interest lies in dynamic scenes, where deformable tracking has received less attention. This is precisely the problem we tackle in this paper; we introduce PD^{2}T, a meticulously studied adaptive deformable tracking pipeline. Such an architecture is invaluable in case the tracking outcome is used for training algorithms, e.g. [8], [9], [10].

Deformable tracking ([11], [12], [13], [14]) aims at tracking the shape of a deformable object in a sequence of frames[1]. This shape typically consists of a number of landmark (i.e. fiducial) points, while the principal 'object' to date is the human face. Deformable tracking methods are separated into i) generic and ii) adaptive. The adaptive methods, often alleged person-specific methods, learn the statistics from (previous) frames and fit the fiducial points with these statistics. On the contrary, in the generic methods a pre-trained landmark localisation technique, typically trained in thousands of images, is employed to fit each frame independently with generic statistics. In this work, we argue that generic methods are suboptimal for tracking, since they fail to capture any correlation among sequential frames and we demonstrate that any generic method could be greatly boosted by employing our adaptive method on the generic results.

The typical way to perform generic deformable tracking, as described in the recent review of [15], consists in a two-step process of i) obtaining the bounding box, ii) performing landmark localisation per frame. The first step of obtaining the bounding box can either be performed by a pre-trained object detector/tracker, or by a model-free tracker. A shortcoming of the generic detection is that small intensity differences between two sequential frames might result in non-overlapping detected regions of interest, due to the (implicit) ranking performed by detectors. An alternative to generic object detection is the model-free tracking techniques. Given the state (typically a rectangular bounding box) of the first frame, the technique estimates the state of the subsequent frames, while no prior information about the object is provided. The tracker should adapt to any appearance/shape changes of the object. Tracking an arbitrary object in an arbitrary video, often alleged 'in-the-wild', constitutes a very challenging task, thus an immense amount of diverse techniques has been proposed [16], [17], [18]. The two major drawbacks of the model-free trackers are that i) they rely on a markov assumption, i.e. there should be no cut frames/change of camera, ii) they are prone to drifting.

In contrast to the generic models, adaptive deformable facial models have received less attention, which can be partially attributed to their computational complexity. Previous lines of research in adaptive deformable tracking have focused on performing joint alignment of the frames given the initial landmarks' estimates. The majority of the works are based on RASL [19] which assumes that a collection of M sequential frames of an object, e.g. a human face, can be decomposed into a low-rank matrix and a sparse error matrix. The drawback of this work is that it allows the object to deform arbitrarily, often leading to unnatural deformations. To mitigate that, Cheng et al. [20] modify RASL by assuming that there are some generic anchor shapes and penalise deformations from anchor shapes. This forces the solution to be close to the initial erroneous fitting. In RAPS [21], they tackle the unnatural deformations by restricting the object to be in an appropriate subspace learnt from clean static images of that object. A significant drawback with these approaches is that they are designed for batches of images and not lengthy clips, in which it is very computationally demanding to fit all images simultaneously. Additionally, the low-rank assumption is restrictive in lengthy clips with extensive object movement. The global SDM of [11] only utilises the temporal correlation to initialise the fitting from the outcome of the previous frame, however this does not i) learn any statistics from the clip, ii) avoid drifting issues.

- *G. Chrysos and S. Zafeiriou are with the Department of Computing, Imperial College London, SW7 2AZ, UK.*
e-mails: {g.chrysos, s.zafeiriou}@imperial.ac.uk

[1] In this paper we use interchangeably the terms sequence of frames, clip and video.

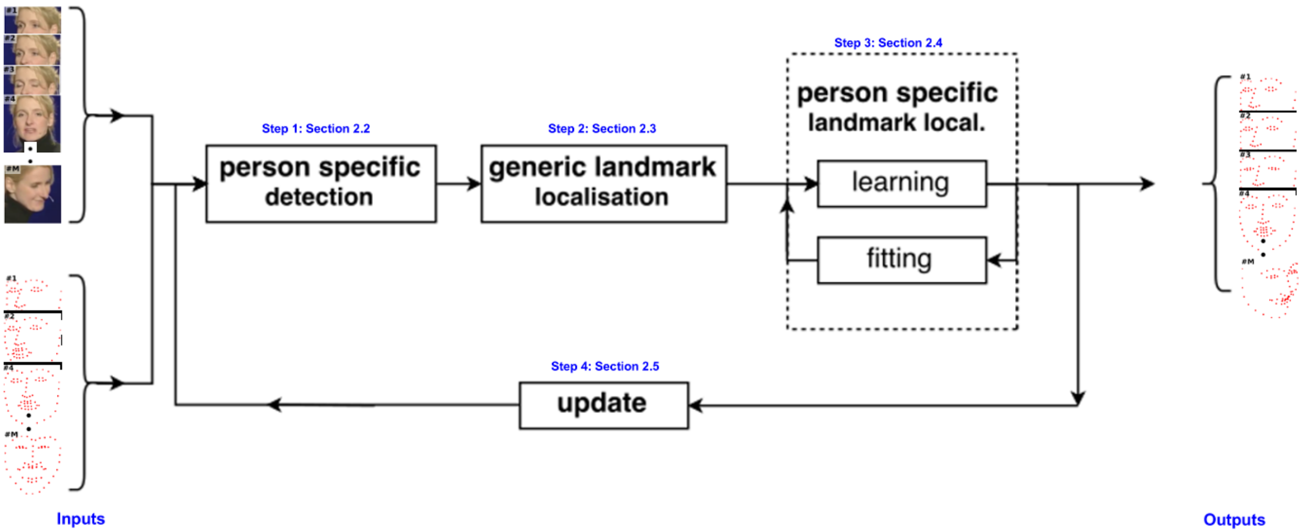


Fig. 1: Visual overview of PD^{2}T. PD^{2}T accepts a sequence of frames and a (not necessarily dense temporally) sequence of landmarks as input; it learns person-specific representations for the object of interest and it outputs an improved sequence of landmarks. There are three main steps in PD^{2}T followed by an optional iterative procedure.

A special part of adaptive models are the incremental learning techniques ([22], [23], [24]). Those are used when online performance is of paramount significance. Such methods are based on a pre-trained model which they update online as new samples emerge. The drawbacks of ([22], [23], [24]) are twofold: a) there is not theoretical guarantee that these discriminative methods converge, b) they often perform well, but their fitting empirically converges only reasonably close to the solution. In addition, the methods of [23], [24] are based on simulation statistics, which are computationally expensive.

An additional reason for lack of adaptive methods was the fact that there was no comprehensive benchmark for evaluating facial landmark tracking methodologies. Evaluation was mainly conducted by simple visual inspection of cherry picked videos [14], [25]. The first large in-the-wild benchmark [9], alleged 300vW, consists of 114 videos (64 for testing) with varying degrees of difficulty. Each video contains a single face, while the annotation is dense in the temporal domain with a sparse set of 68 fiducial points tracked. Annotating hundreds of thousands of frames (especially in the challenging category 3) was a gargantuan task. Consequently, we had to devise methodologies that automatically provide an accurate landmark localisation output to minimise the manual annotation effort. As we have demonstrated in the past [26] methodologies similar to this work aid in large scale video annotation.

In this work, we support that object alignment per frame is suboptimal; much richer representations can be extracted from sequential frames. To that end, we introduce PD^{2}T, a fully-automatic pipeline, that considers as input a number of sequential frames along with a generic fitting per frame, learns person-specific statistics and fits the learnt models to the frames. The proposed pipeline is a 3-step process where 1) the initial step is used to learn person-specific statistics for improved detection, then 2) using any off-the-shelf generic landmark localisation as a second step, and finally 3) learning a person-specific model for refining the landmark positions per frame. This process can be iterated several times which will iteratively improve the final deformable tracking results[2]. We propose two versions of PD^{2}T: i) the incremental (online), ii) the offline one. The incremental version learns the models from the first few frames that are available (100 in the experiments), then the learnt models are updated incrementally; the offline version accepts as input the complete sequence/set of generic results. The offline version allows the relaxation of the markov assumption, since there is both a detection and a localisation step in each frame. The motivation behind the offline version is the optimal performance for demanding applications, while the incremental version offers the chance of online processing. Thorough experimentation is performed in both human faces and animal tracking to validate PD^{2}T's effectiveness. The generic methods of [15] are considered as input in 300vW and demonstrate that PD^{2}T improves the outcomes in **every single case**. Due to the saturation of top-performing methods in 300vW, we introduce FTOVM, a new dataset with harder cases that have not emerged in 300vW. Additionally, a number of videos are annotated in a previously unstudied task, i.e. animal landmark tracking, in order to investigate the robustness of our method to less established objects. The animal videos will be publicly available in https://ibug.doc.ic.ac.uk/.

A preliminary version of this work has appeared in [26], however a number of extensions are performed with most prominent the fact that PD^{2}T can be used for several deformable objects; experiments with both human facial tracking and animal tracking are provided. Additionally, in this work we present an incremental (online) version of the pipeline, which includes incremental learning of statistics for both detection and landmark localisation. This has not emerged in the past, while the incremental pictorial structures (person-specific detection) have not emerged in the literature before. Moreover, the objective of this work consists in learning an adaptive model on top of any generic deformable tracking method. A third difference is the extensive self evaluation part includes experimentation over the architectural/algorithmic decisions to optimise the pipeline.

[2] In practice we found that one iteration is sufficient in all our experiments, i.e. the person-specific detection and then landmark localisation were performed only once.

Our contributions are summarised as follows:

- An adaptive object-agnostic pipeline, which we name PD^{2}T, is proposed. We propose an i) incremental, ii) offline version of PD^{2}T with several self evaluation experiments (deferred to the supplementary material). This pipeline ameliorates the main drawbacks of the generic methods.
- The incremental version of the pipeline includes the incremental pictorial structures for detection, which has never emerged before.
- Experimentation in an ad-hoc scenario (animal tracking) which have never been published before. We illustrate that PD^{2}T surpasses the state-of-the-art results in both human face and animal tracking.
- Two new datasets are proposed. The FTOVM addresses cases that do not emerge in 300vW, e.g. the out of scene movement. The second dataset, includes cats' faces and was developed for assessing the robustness of PD^{2}T in an ad-hoc deformable tracking scenario.

Even though the current implementation is not real-time, we strongly believe that the proposed approach is extremely valuable; it can provide very accurate large scale landmark annotations which can be used for various tasks (e.g. training deep learning methodologies, organising benchmarks).

The rest of the paper is organised as follows: In Sec. 2 the proposed method is developed, the algorithms for every steps are introduced; in Sec. 3 the experimental comparisons and results are described.

2 Method

Given i) a set of M sequential frames $\boldsymbol{D} = \{\boldsymbol{i}^{(1)}, \boldsymbol{i}^{(2)}, ..., \boldsymbol{i}^{(M)}\}$ along with ii) an initial set of landmarks in a subset of the frames, the objective of PD^{2}T is to improve the initial deformable tracking results by learning person-specific models. To achieve the optimal localisation outcomes, the two core steps of PD^{2}T are alternated: improve the bounding box detection, improve the landmark localisation. Each step of the pipeline is analysed in the subsequent Sections, while a block diagram of the pipeline is visualised in Fig. 1. The fundamental assumption of our method is that there are at least few frames in each video that are initially well fitted. Experimentally, we verify that approximately 50 frames suffice for a clip over a minute long ($\sim$ 1800 frames), which should be achievable by several existing localisation techniques.

2.1 Notation

A capital (small) bold letter represents a matrix (vector), while a plain letter designates a scalar number. The following notational simplifications are performed in the succeeding Sections: (a) The time t is dropped if unnecessary, e.g. instead of writing frame $\boldsymbol{i}^{(t)}$, it will be expressed as $\boldsymbol{i}$, (b) an 'image patch' refers by default to the vectorised version of the patch, (c) the feature extraction function $\boldsymbol{\phi}$ is implicitly assumed in patches, i.e. when denoting a patch as $\boldsymbol{i}_j$, it refers to $\boldsymbol{\phi}(\boldsymbol{i}_j)$.

The purpose of landmark tracking in each frame $\boldsymbol{i}$ consists in localising a set of n points with configuration

$$\boldsymbol{l} = [[\boldsymbol{\ell}_1]^T, [\boldsymbol{\ell}_2]^T, \dots, [\boldsymbol{\ell}_n]^T]^T = [x_1, y_1, x_2, y_2, \dots, x_n, y_n]^T$$

with $\boldsymbol{\ell}_j = [x_j, y_j]^T, j \in [1, n]$ the Cartesian coordinates of the j$^{\text{th}}$ point in time t. An image patch centered around the point $\boldsymbol{\ell}_j$ is denoted as $\boldsymbol{i}_j$. If the patch shape is (p_w, p_h) and the patch area is defined as $p_a = p_w \cdot p_h$, then $\boldsymbol{i}_j \in \mathbb{R}^{p_a}$. The relative location of two points of interest is determined as the vector of their spatial difference, i.e. $\boldsymbol{\ell}_j - \boldsymbol{\ell}_k = [x_j - x_k, y_j - y_k]^T$. The $\boldsymbol{\mathcal{I}}$ denotes an identity matrix of appropriate dimensionality. The primary symbols in the manuscript are summarised in Tab. 1.

TABLE 1: Summary of primary symbols.

Symbol	Dimension(s)	Definition
n	$\mathbb{R}$	Number of landmark points in a frame.
M	$\mathbb{R}$	Number of frames in the video.
$\boldsymbol{\ell}_j$	$\mathbb{R}^2$	Cartesian coordinates of the j$^{\text{th}}$ landmark point.
$\boldsymbol{l}$	$\mathbb{R}^{2n}$	Spatial configuration for all landmarks.
$\boldsymbol{i}$	$\mathbb{R}^{h \cdot w}$	Vectorised frame (image).
$\boldsymbol{i}_j$	$\mathbb{R}^{p_a}$	Image patch around $\boldsymbol{\ell}_j$ with area p_a.
$\boldsymbol{\mathcal{I}}$	-	Identity matrix of appropriate dimensionality.

2.2 Person-specific pictorial detection

The first phase of PD^{2}T consists in performing a more accurate detection by utilising the pictorial structures of [27]. A generative method is preferred, since empirically generative methods require less frames than their discrimative counterparts to learn object representations; an experimental validation of this claim is performed in the self evaluation part. For a frame $\boldsymbol{i}$, we want to confidently locate the position $\boldsymbol{l}_j$ that is very close to the ground-truth position $\boldsymbol{l}_j^{gt}$ for all points j. We express that with the joint probability $P(\boldsymbol{i}, \boldsymbol{l}|\boldsymbol{A}, \boldsymbol{S})$ where $\boldsymbol{A}, \boldsymbol{S}$ convey the appearance and deformation parameters respectively.

A tree $T = (V, E)$ is constructed online for utilising the efficient derivation conducted by Felzenszwalb *et al.*in [27], [28]. Each vertex $V = \{v_1, v_2, ..., v_n\}$ corresponds to the texture of a point of interest j, while each edge E models the spatial constraints between every pair of points that are connected. The optimal configuration for the edges E is provided by computing the Minimum Spanning Tree (MST), which consitutes a tree with minimum total weights on the edges. To learn the set of edges, a complete graph with the vertices V and the edges is initialised and then Kruskal's algorithm is applied to compute the MST, please consult [27] for further details.

With the typical assumption that the patch appearance is independent from other patches (frequently appearing in computer vision, e.g. in [2], [27], [29]) and the restricted format of the tree G, the joint probability is

$$\begin{gathered} P(\boldsymbol{i}, \boldsymbol{l}|\boldsymbol{C}) = P(\boldsymbol{i}|\boldsymbol{l}, \boldsymbol{A})P(\boldsymbol{l}|\boldsymbol{S}) = \\ \prod_{j=1}^{n} P(\boldsymbol{i}_j|\boldsymbol{\ell}_j, \boldsymbol{A}) \prod_{(v_k, v_j) \in E}^{n} P(\boldsymbol{\ell}_j, \boldsymbol{\ell}_k|\boldsymbol{S}) \end{gathered} \tag{1}$$

We want to maximise this probability or correspondingly minimise the probability's negative logarithm. The first term of Eq. 1 corresponds to the appearance term, while the second to the spatial configuration of the points. We model each of the two terms with a multivariate Gaussian distribution, which allows us to use the generalised distance transforms of [28] for computing the cost per frame. Each term is developed independently below and the combined cost function is formulated.

2.2.1 *Appearance modelling*

A multivariate Gaussian distribution $\mathcal{N}(\boldsymbol{i}_j|\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)$ with $\boldsymbol{\mu}_j \in \mathbb{R}^{p_a}$ the mean representation, $\boldsymbol{\Sigma}_j \in \mathbb{R}^{p_a \times p_a}$ the covariance, represents the appearance of each patch $\boldsymbol{i}_j$. By considering the negative logarithm of the appearance term of Eq. 1, the optimisation is reduced to searching for the patch $\boldsymbol{i}_j$ that minimises the Mahalanobis distance. Mathematically,

$$\arg\min_{\boldsymbol{i}_j}(\boldsymbol{i}_j-\boldsymbol{\mu}_j)^T\boldsymbol{\Sigma}_j^{-1}(\boldsymbol{i}_j-\boldsymbol{\mu}_j) \tag{2}$$

A Singular Value Decomposition (SVD) is performed on $\boldsymbol{\Sigma}_j \approx \boldsymbol{U}_j\boldsymbol{\mathcal{L}}_j\boldsymbol{U}_j^T$ with $\boldsymbol{U}_j \in \mathbb{R}^{p_a\times m}$, $\boldsymbol{\mathcal{L}}_j \in \mathbb{R}^{m\times m}$ to reduce the variance. However, we have experimentally noticed that the hard cases are not covered by the Gaussian distribution assumption, hence we augment our formulation by assuming that there is a latent representation that follows the Gaussian distribution. Each patch $\boldsymbol{i}_j$ is modelled as a linear combination of three terms: (a) a projection of a random variable $\boldsymbol{r}_j$ to $\boldsymbol{U}_j$, (b) the mean appearance $\boldsymbol{\mu}_j$ and (c) a random variable $\boldsymbol{\varepsilon}$ as the noise term. Equivalently in math formulation:

$$\boldsymbol{i}_j = \boldsymbol{U}_j\boldsymbol{r}_j + \boldsymbol{\mu}_j + \boldsymbol{\varepsilon} \tag{3}$$

where $\boldsymbol{r}_j \sim \mathcal{N}(\boldsymbol{r}_j|\boldsymbol{0},\boldsymbol{\mathcal{L}_j})$, $\boldsymbol{\varepsilon} \sim \mathcal{N}(\boldsymbol{\varepsilon}|\boldsymbol{0},\sigma^2\boldsymbol{\mathcal{I}})$. Then the marginal distribution of $\boldsymbol{i}_j$ is a Gaussian formulated as:

$$\begin{aligned}P(\boldsymbol{i}_j|\boldsymbol{\ell}_j,\boldsymbol{A}) = \int_{\boldsymbol{r}_j} P(\boldsymbol{i}_j|\boldsymbol{r}_j,\boldsymbol{\ell}_j,\boldsymbol{A})P(\boldsymbol{r}_j|\boldsymbol{A})\boldsymbol{dr}_j = \\ \mathcal{N}(\boldsymbol{i}_j|\boldsymbol{\mu}_j,\boldsymbol{\Sigma}_j^{augm})\end{aligned} \tag{4}$$

with $\boldsymbol{\Sigma}_j^{augm} = \boldsymbol{U}_j\boldsymbol{\mathcal{L}}_j\boldsymbol{U}_j^T + \sigma^2\boldsymbol{\mathcal{I}}$, $\boldsymbol{\Sigma}_j^{augm} \in \mathbb{R}^{p_a\times p_a}$. Applying the Woodbury formula on the augmented sigma, we obtain

$$(\boldsymbol{\Sigma}_j^{augm})^{-1} = \boldsymbol{U}_j(\boldsymbol{\mathcal{L}}_j+\sigma^2\boldsymbol{\mathcal{I}})^{-1}\boldsymbol{U}_j^T + \frac{1}{\sigma^2}(\boldsymbol{\mathcal{I}}-\boldsymbol{U}_j\boldsymbol{U}_j^T) \tag{5}$$

The latent representation of Eq. 3 led to a Gaussian representation (Eq. 4) of the patch $\boldsymbol{i}_j$ with the initial mean and an augmented variance ($\boldsymbol{\Sigma}_j^{augm}$), thus as derived above we optimise for

$$\begin{aligned}\arg\min_{\boldsymbol{i}_j}(\boldsymbol{i}_j-\boldsymbol{\mu}_j)^T(\boldsymbol{\Sigma}_j^{augm})^{-1}(\boldsymbol{i}_j-\boldsymbol{\mu}_j) \overset{Eq.5}{=} \\ \arg\min_{\boldsymbol{i}_j}[(\boldsymbol{i}_j-\boldsymbol{\mu}_j)^T\boldsymbol{U}_j(\boldsymbol{\mathcal{L}}_j+\sigma^2\boldsymbol{\mathcal{I}})^{-1}\boldsymbol{U}_j^T(\boldsymbol{i}_j-\boldsymbol{\mu}_j)+ \\ (\boldsymbol{i}_j-\boldsymbol{\mu}_j)^T\frac{1}{\sigma^2}(\boldsymbol{\mathcal{I}}-\boldsymbol{U}_j\boldsymbol{U}_j^T)(\boldsymbol{i}_j-\boldsymbol{\mu}_j)]\end{aligned} \tag{6}$$

In the last equation there are two parts: (a) the Mahalanobis distance within the subspace $\boldsymbol{U}_j$, (b) the distance to its orthogonal supplement ($\boldsymbol{\mathcal{I}}-\boldsymbol{U}_j\boldsymbol{U}_j^T$) scaled by the inverse variance term σ^2. The latter distance term captures appearances that cannot be reconstructed from the learned subspace $\boldsymbol{U}_j$; by combining both terms we model more effectively the whole texture space.

2.2.2 Spatial Modelling

The relative position $(\boldsymbol{\ell}_j-\boldsymbol{\ell}_k)$ of a vertex (point of interest) j from its parent vertex k is modelled as a multivariate Gaussian distribution $\mathcal{N}((\boldsymbol{\ell}_j-\boldsymbol{\ell}_k)|\boldsymbol{\mu}_{j,k},\boldsymbol{\Sigma}_{j,k})$ with $\boldsymbol{\mu}_{j,k} \in \mathbb{R}^2$ the mean and $\boldsymbol{\Sigma}_{j,k} \in \mathbb{R}^{2\times 2}$ the covariance. Then the negative logarithm of $P(\boldsymbol{\ell}_j,\boldsymbol{\ell}_k|\boldsymbol{\mu}_{j,k},\boldsymbol{\Sigma}_{j,k},E)$ from Eq. 1 that corresponds to the deformation is optimised by

$$\arg\min_{\boldsymbol{\ell}_j,\boldsymbol{\ell}_k}((\boldsymbol{\ell}_j-\boldsymbol{\ell}_k)-\boldsymbol{\mu}_{j,k})^T\boldsymbol{\Sigma}_{j,k}^{-1}((\boldsymbol{\ell}_j-\boldsymbol{\ell}_k)-\boldsymbol{\mu}_{j,k}) \tag{7}$$

This optimisation is computationally demanding as there are thousands of candidate positions in a 2D grid while the computation should consider the combinations of the $\boldsymbol{\ell}_j$ and the parent position of $\boldsymbol{\ell}_k$; a naive implementation would require quadratic time in the number of positions in the grid. Utilising the tree structure and the generalised distance transform, the computation is performed in linear time. Additionally, in this work only the diagonal elements of the covariance matrix $\boldsymbol{\Sigma}_{j,k}$ are considered in the computation of the aforementioned optimisation.

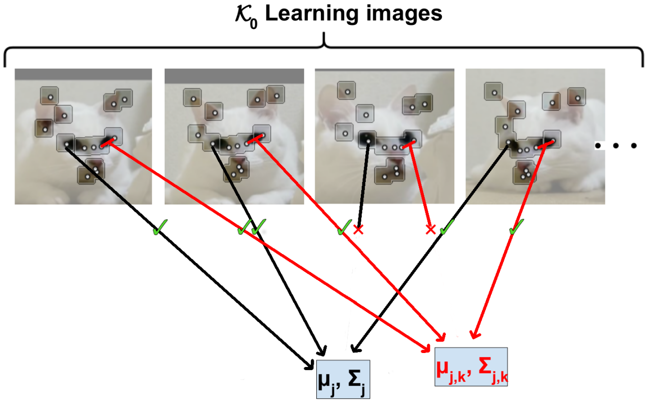


Fig. 2: (Preferably viewed in colour) Pictorial representation of the person-specific detection learning. From top to bottom, the steps of a) extracting the patches from the current landmarks' estimates, b) rejecting the erroneous fittings with the classifier, c) learning the appearance (unary) and spatial (pairwise) subspaces.

2.2.3 Learning

The pictorial representation parameters are estimated from the input, which is a subset $\boldsymbol{D}_p$ ($\boldsymbol{D}_p \subseteq \boldsymbol{D}$) of (not necessarily sequential) frames along with the respective spatial configuration. Both the appearance parameters $\boldsymbol{A} = \{\boldsymbol{\mu}_j,\boldsymbol{\Sigma}_j|\forall j, j \in [1,n]\}$ and the set $\{\boldsymbol{\mu}_{j,k},\boldsymbol{\Sigma}_{j,k}|\forall(j,k),(v_k,v_j) \in E\}$ from the deformation parameters are learnt with maximum likelihood.

Due to the sensitivity of the covariances to outlier fittings, only well-fitted frames/landmarks pairs are considered as learning samples. For this purpose, a classifier is applied to reject the erroneous fittings, i.e. a function $f_{cl}(\boldsymbol{i},\boldsymbol{l})$ that accepts a frame $\boldsymbol{i}$ along with the respective fitting $\boldsymbol{l}$ and returns a binary decision on whether this is an acceptable fitting. In our case, f_{cl} represents a linear patch-based SVM of [30]. A set of $\mathcal{K}_0$ accepted fittings is used to learn the pictorial representation parameters; the learning procedure is visually depicted in Fig. 2.

2.2.4 Incremental pictorial detection

In the online version of PD^{2}T, the pictorial detection is learnt in the first few frames, hence adaptations of the parameters (means, covariances) are required over time. In this Section the updates of the parameters are developed (only applicable to the online version). The incremental update in our case can be exact ([31]), i.e. it is equivalent to training a new appearance model with an augmented set of images $[\boldsymbol{i}^{(\mathcal{K}_0+1)},\ldots,\boldsymbol{i}^{(\mathcal{K}_{new})}]$. The update of the appearance parameters is described below, while for the deformation parameters it follows a similar logic. In order to avoid defining new auxiliary variables, the derivation below will assume momentarily that $\boldsymbol{\Sigma}_j^{augm} = \boldsymbol{U}_j\boldsymbol{\mathcal{L}}_j\boldsymbol{U}_j^T$, however exactly the same derivation can be made if an SVD is performed in the full $\boldsymbol{\Sigma}_j^{augm}$.

Given the previous mean $\boldsymbol{\mu}_j$, the subspace $\boldsymbol{U}_j$, the diagonal matrix $\boldsymbol{\mathcal{L}}_j$ and the new data $\boldsymbol{B} = [\boldsymbol{i}^{(\mathcal{K}_0+1)},\ldots,\boldsymbol{i}^{(\mathcal{K}_{new})}]$ the goal of the update consists in learning a $\hat{\boldsymbol{\mu}}_j$ and a $\hat{\boldsymbol{U}}_j$, $\hat{\boldsymbol{\mathcal{L}}}_j$ with

$\hat{\boldsymbol{\Sigma}}_j = \hat{\boldsymbol{U}}_j \, \hat{\boldsymbol{\mathcal{L}}}_j \, \hat{\boldsymbol{U}}_j^T$. The $\mathcal{K}_{new}$ samples can be expressed as a linear combination of the components already included in $\boldsymbol{U}_j$ and the components to the orthogonal subspace to $\boldsymbol{U}_j$ (denoted as $\boldsymbol{V}_j$). Then the updated sigma $\hat{\boldsymbol{\Sigma}}_j$ is equal to

$$\hat{\boldsymbol{\Sigma}}_j = [\boldsymbol{U}_j \ \boldsymbol{V}_j] \begin{bmatrix} \boldsymbol{\mathcal{L}}_j & \boldsymbol{U}_j^T \boldsymbol{B} \\ \boldsymbol{0} & \boldsymbol{V}_j^T \boldsymbol{B} \end{bmatrix} \begin{bmatrix} \boldsymbol{U}_j^T & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{\mathcal{I}} \end{bmatrix} \tag{8}$$

An SVD is performed in the matrix $\begin{bmatrix} \boldsymbol{\mathcal{L}}_j & \boldsymbol{U}_j^T \boldsymbol{B} \\ \boldsymbol{0} & \boldsymbol{V}_j^T \boldsymbol{B} \end{bmatrix} = \tilde{\boldsymbol{U}}_j \, \tilde{\boldsymbol{\mathcal{L}}}_j \, \tilde{\boldsymbol{U}}_j^T$, which allows us to rewrite Eq.8 as

$$\hat{\boldsymbol{\Sigma}}_j = ([\boldsymbol{U}_j \ \boldsymbol{V}_j] \, \tilde{\boldsymbol{U}}_j) \, \tilde{\boldsymbol{\mathcal{L}}}_j \, (\tilde{\boldsymbol{U}}_j^T \begin{bmatrix} \boldsymbol{U}_j^T & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{\mathcal{I}} \end{bmatrix}) \tag{9}$$

The latter expression of $\hat{\boldsymbol{\Sigma}}_j$ enables the definition of the updated terms as $\hat{\boldsymbol{U}}_j = [\boldsymbol{U}_j \ \boldsymbol{V}_j] \ \tilde{\boldsymbol{U}}_j$ and $\hat{\boldsymbol{\mathcal{L}}}_j = \tilde{\boldsymbol{\mathcal{L}}}_j$, while the $\hat{\boldsymbol{\mu}}_j$ is the weighted mean of $\boldsymbol{\mu}_j$ combined with the mean of the new image patches, which concludes the update rules.

2.2.5 Cost function

By combining the Eq. $\{4, 5, 7\}$, the final cost function is derived from Eq. 1 as

$$\begin{aligned} \underset{\boldsymbol{i}^{all}, \boldsymbol{l}}{\arg\min} \sum_{j=1}^{n} [(\boldsymbol{i}_j - \boldsymbol{\mu}_j)^T \boldsymbol{U}_j (\boldsymbol{\mathcal{L}}_j + \sigma^2 \boldsymbol{\mathcal{I}})^{-1} \boldsymbol{U}_j^T (\boldsymbol{i}_j - \boldsymbol{\mu}_j) + \\ (\boldsymbol{i}_j - \boldsymbol{\mu}_j)^T \frac{1}{\sigma^2} (\boldsymbol{\mathcal{I}} - \boldsymbol{U}_j \boldsymbol{U}_j^T)(\boldsymbol{i}_j - \boldsymbol{\mu}_j) + \\ ((\boldsymbol{\ell}_j - \boldsymbol{\ell}_k) - \boldsymbol{\mu}_{j,k})^T \boldsymbol{\Sigma}_{j,k}^{-1} ((\boldsymbol{\ell}_j - \boldsymbol{\ell}_k) - \boldsymbol{\mu}_{j,k})] \end{aligned} \tag{10}$$

with $\boldsymbol{l} = [[\boldsymbol{\ell}_1]^T, [\boldsymbol{\ell}_2]^T, \ldots, [\boldsymbol{\ell}_n]^T]^T$ the spatial configuration and $\boldsymbol{i}^{all} = [\boldsymbol{i}_1^T, \boldsymbol{i}_2^T, \ldots, \boldsymbol{i}_n^T]^T$ the respective patches for all the parts. For each part j (vertex in T) there are three terms: (a) the unary cost computed as the Mahalanobis distance within the subspace $\boldsymbol{U}_j$, (b) the unary cost from the orthogonal subspace for complementary appearance information, (c) the deformation from the nominal displacement from its' parent position (k denotes the parent node of vertex j).

2.2.6 Efficient inference

Note in Eq. 10 that the total cost of each part j is defined with respect to its parent's vertex position, while the sum of all parts' costs should be minimised. This dependency does not allow us to independently perform the score computation per part, however by utilising a message passing scheme, we can compute all the costs in a single forward pass (from the leaves to the root) and then backtrack to locate the patches and the parts.

Starting from the leaf nodes and inversely traversing the tree towards the root vertex, the cost of placing each vertex in every position of the grid is computed by utilising the generalised distance transforms. This cost is passed as a message to the parent of the vertex, which performs a similar cost computation till all the costs are summed in the root vertex. The optimal position $\boldsymbol{\ell}_{root}$ is computed and then backtrack utilising the lookup tables to locate the position $\boldsymbol{\ell}_j$ of each part j, hence the patches and the spatial configuration are decided.

2.3 Generic landmark localisation

From the pictorial step, we obtain for every frame $\boldsymbol{i}^{(t)} \in \boldsymbol{D}$ a spatial configuration $\boldsymbol{l}^{(t)}$. The tightest bounding box of $\boldsymbol{l}^{(t)}$ is computed and a generic landmark localisation technique is applied in this step.

The recent development of such statistical methods can be divided into two major categories: (a) *discriminative* models that employ regression, e.g. [5], [7], [25], and (b) *generative* models, e.g. [32], [33], [34]. As recent benchmarks have illustrated, the discriminative methods with a cascade of regressors [5] outperform the rest methods. In our case, any standard landmark localisation technique can be applied, either the same as the one applied as the pre-pipeline initialisation or any other off-the-shelf method (including the discriminative state-of-the-art methods). In our experiments we consistently use the one that was applied as the initialisation before the call of the pipeline. This step is required as an initialisation to the person specific localisation method of the following step.

2.4 Person-specific landmark localisation

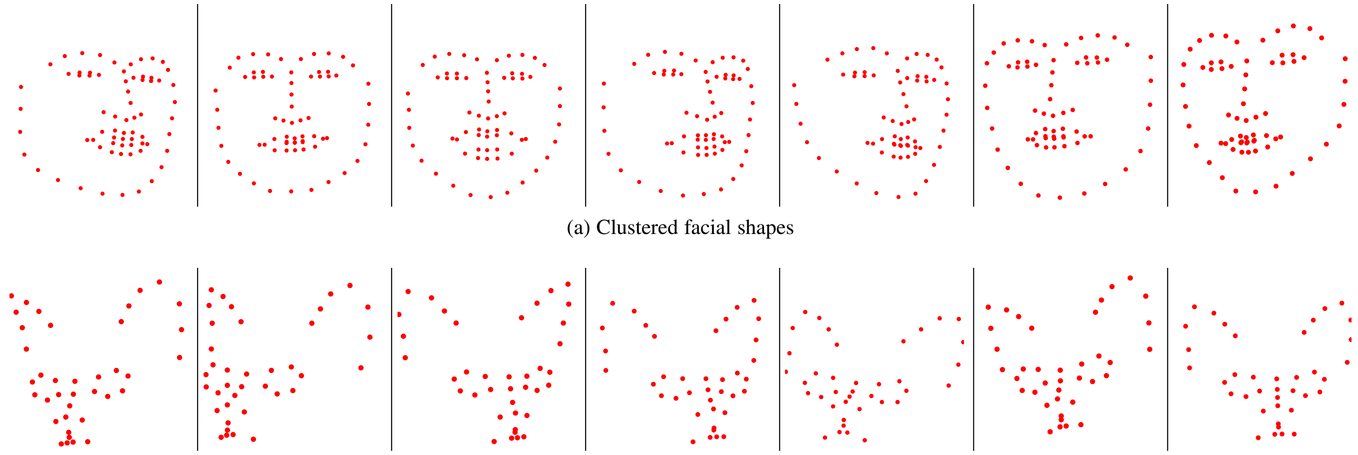
In this step we aim to refine the generic fittings of the previous step by constructing a model that best describes the sequence's variation. We choose the generative Active Appearance Models of [32] that have empirically demonstrated great results when the initialisation is quite close to the ground-truth. Specifically, we employ the state-of-the-art part-based Active Appearance Model (AAM) of [34], referred to as Gauss-Newton DPM (GN-DPM), and iteratively update the appearance model to improve the fitting results.

GN-DPM of [34] is a generative statistical model of shape and appearance that recovers a parametric description of an object through Gauss-Newton optimization. Generalized Procrustes Analysis is utilised to align a set of K configurations $\{\boldsymbol{l}^{(1)}, \ldots, \boldsymbol{l}^{(K)}\}$; Principal Component Analysis (PCA) is utilised to learn an orthonormal basis of n_l eigenvectors $\boldsymbol{U}^{(l)} \in \mathbb{R}^{2n \times n_l}$ and the mean shape $\bar{\boldsymbol{l}}$. This linear shape model can be used to generate a shape instance as $\boldsymbol{l}(\boldsymbol{p}) = \bar{\boldsymbol{l}} + \boldsymbol{U}^{(l)} \boldsymbol{p}$, where $\boldsymbol{p} = [p_1, \ldots, p_{n_l}]^T$ is the vector of shape parameters. The appearance model is constructed in a similar way; a patch based representation is extracted per part; the representation is converted into a feature vector, i.e. $\boldsymbol{i}_j$ for landmark location j in each frame $\boldsymbol{i}$; the feature vectors are concatenated to form a single vectorized part-based appearance representation. A PCA is utilised in the set of part-based appearance vectors that results in a subspace of n_A eigenvectors $\boldsymbol{U}^{(A)} \in \mathbb{R}^{(n \cdot b) \times n_A}$ (b is the feature length of a vectorised featured patch) and the mean appearance $\bar{\boldsymbol{a}}$. This model can be used to synthesize shape-free appearance instances, as $\boldsymbol{a}(\boldsymbol{\lambda}) = \bar{\boldsymbol{a}} + \boldsymbol{U}^{(A)} \boldsymbol{\lambda}$, where $\boldsymbol{\lambda} = [\lambda_1, \ldots, \lambda_{n_A}]^T$ is the vector of appearance parameters. Given a test image $\mathbf{I}$, the optimization problem of GN-DPM employs an ℓ_2^2 norm and is expressed as

$$\begin{aligned} \underset{\boldsymbol{p}, \boldsymbol{\lambda}}{\arg\min} \, \| \boldsymbol{a}(\boldsymbol{l}(\boldsymbol{p}) | \boldsymbol{I}) - \boldsymbol{a}(\boldsymbol{\lambda}) \|^2 = \\ \underset{\boldsymbol{p}, \boldsymbol{\lambda}}{\arg\min} \, \| \boldsymbol{a}(\bar{\boldsymbol{l}} + \boldsymbol{U}^{(l)} \boldsymbol{p} | \boldsymbol{I}) - \boldsymbol{a}(\boldsymbol{\lambda}) \|^2 \end{aligned} \tag{11}$$

This can be efficiently solved using the Gauss-Newton algorithm with a detailed derivation in [34].

In our case, we formulate an iterative optimization problem that aims to minimize the mean GN-DPM fitting ℓ_2 norm of Eq. 11 over all the M frames of the video. As a generative



(a) Clustered facial shapes

(b) Clustered cats' facial shapes

Fig. 3: Indicative shapes for each cluster. Even with the seven clusters visualised in the figure significant shape variance is included. This ensures that the shape model (as detailed in Sec. 2.4) contains considerable variance.

method GN-DPM requires few well chosen images for learning appropriate appearance/shape subspaces with maximum variance. A two step process is performed for selecting the samples for learning. Initially, the pool of samples is reduced by rejecting the erroneous fittings; sequentially we perform clustering and draw the samples from each cluster. The rejection of the erroneous fittings is performed with the same classifier as in the pictorial learning (Sec. 2.2.3), i.e. all the configurations $\{\boldsymbol{l}^{(1)}, \ldots, \boldsymbol{l}^{(M)}\}$ along with the respective patches are inserted into the function $f_{cl}(\boldsymbol{i}, \boldsymbol{l})$, which outputs a number of accepted configurations/frames. Only a subset of the accepted configurations is used to form the person-specific shape subspace $\boldsymbol{U}^{(l)}$ and the person-specific appearance subspace $\boldsymbol{U}^{(A)}$. To achieve the maximum variance in the shape representation, clustering is performed in the shape coordinates and a single instance is drawn from each cluster. All the accepted shapes are aligned with Procrustes; a Gaussian Mixture Model (GMM) with a predefined number of components is fit; one sample is drawn from each cluster. Each shape along with the respective texture is mirrored horizontally; this aggregated set consists the set of learning samples for the subspaces $\boldsymbol{U}^{(l)}$ and $\boldsymbol{U}^{(A)}$. In Fig.3 few indicatives samples drawn from such a GMM for a clip are visualised. Sequentially, a generic subspace, symbolised as $\boldsymbol{G}^{(l)}$, is formed by augmenting the subspace $\boldsymbol{U}^{(l)}$ with a set of Q generic shapes, i.e. shapes from datasets for static image alignment. As developed in the experimental section less than 10 images are added with the purpose to ensure that some extreme poses are included in the shape variation. The optimisation function of this step is

$$\begin{aligned}&\arg\min_{\bar{\boldsymbol{l}},\bar{\boldsymbol{a}},[\boldsymbol{U}^{(l)}\ \boldsymbol{G}^{(l)}],\boldsymbol{p}^t,\boldsymbol{\lambda}^t} \frac{1}{M}\sum_{t=1}^{M}\|\boldsymbol{a}(\bar{\boldsymbol{l}}+[\boldsymbol{U}^{(l)}\ \boldsymbol{G}^{(l)}]\boldsymbol{p}^t|\boldsymbol{I}^t)-\boldsymbol{a}(\boldsymbol{\lambda}^t)\|^2\\&\text{s.t. } [\boldsymbol{U}^{(l)}\ \boldsymbol{G}^{(l)}]^T[\boldsymbol{U}^{(l)}\ \boldsymbol{G}^{(l)}]=\boldsymbol{\mathcal{I}}\end{aligned} \tag{12}$$

This procedure updates the person-specific shape basis $\boldsymbol{U}^{(l)}$ and ensures that the two shape bases remain orthonormal, i.e. $[\boldsymbol{U}^{(l)}\ \boldsymbol{G}^{(l)}]^T[\boldsymbol{U}^{(l)}\ \boldsymbol{G}^{(l)}]=\boldsymbol{\mathcal{I}}$. Then the parameters $\{\mathbf{p}^t, \boldsymbol{\lambda}^t\}$, $t \in \{1, \ldots, M\}$ are re-estimated by minimizing the ℓ_2 norm for each frame. Thus, the following two steps are performed in an alternating manner:

(1) Fix $\{\mathbf{p}^t, \boldsymbol{\lambda}^t\}$ *and minimize for* $\{\bar{\boldsymbol{\ell}}, \bar{\boldsymbol{a}}, [\boldsymbol{U}^{(l)}\ \boldsymbol{G}^{(l)}]\}$*:* Utilising the current shape estimation for each frame $t \in \{1, \ldots, M\}$ a person-specific shape subspace $\boldsymbol{U}^{(l)}$ is constructed, ensuring that it satisfies the orthogonality constrain with the generic shape subspace.

(2) Fix $\{\bar{\boldsymbol{l}}, \bar{\boldsymbol{a}}, [\boldsymbol{U}^{(l)}\ \boldsymbol{G}^{(l)}]\}$ *and minimize for* $\{\boldsymbol{p}^t, \boldsymbol{\lambda}^t\}$*:* Having created the orthogonal basis that combines the person-specific appearance variation, we now aim to estimate the shape and appearance parameters per frame. Based on Eq. 11, this is achieved by solving

$$\arg\min_{\boldsymbol{p}^t,\boldsymbol{\lambda}^t} \|\boldsymbol{a}(\boldsymbol{l}(\boldsymbol{p}^t)|\boldsymbol{I}^t)-\boldsymbol{a}(\boldsymbol{\lambda}^t)\|^2,\ \forall t \in \{1, \ldots, M\} \tag{13}$$

using the Gauss-Newton optimization [34].

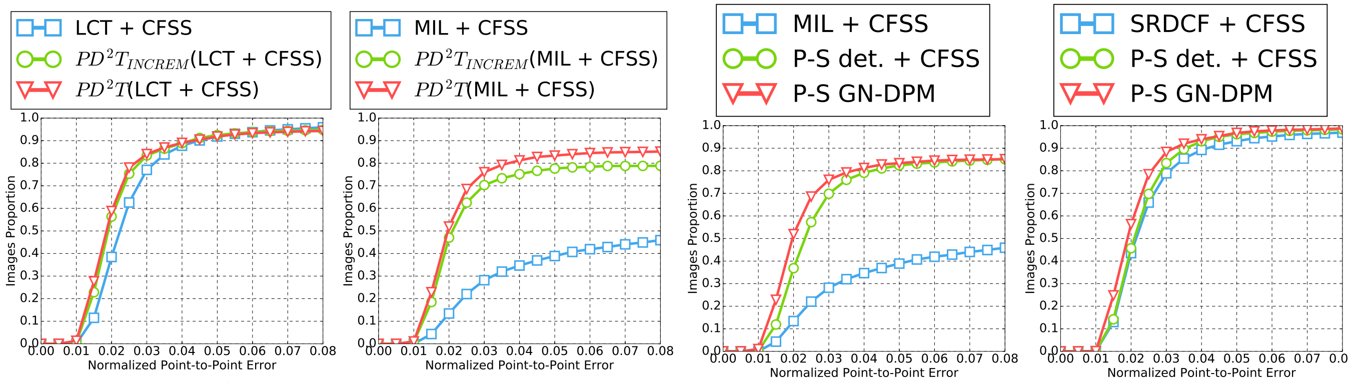
The above iterative procedure gradually improves the statistical model and, as a consequence, the spatial configurations. Our experimentation demonstrated that few iterations suffice to end up with accurate results.

2.5 Update

The learnt statistics of the aforementioned steps improve the tracking outcomes of a generic method, however depending on the nature of the videos and the deformable object, few frames could be further improved by iterating the previous three steps. The configurations $\{\boldsymbol{l}^{(1)}, \ldots, \boldsymbol{l}^{(M)}\}$ obtained from the last step are fed as input to the person-specific detection step; new improved statistics with an initialisation closer to the ground-truth is learnt. In practice, one iteration is sufficient to obtain results on-par with the state-of-the-art deformable tracking results, even if the initial generic method is not very accurate.

3 Experiments

In this Section, we develop the experimental setup and present the experimental results. These are separated into i) deformable facial landmark tracking, ii) ad-hoc deformable object tracking. For deformable facial tracking the generic results of [15] are employed.



(a) PD^{2}T: incremental vs offline version

(b) PD^{2}T internal steps

Fig. 4: (Preferably viewed in colour) CED plots for validation experiments (Sec. 3.2). In all the plots, the blue line indicates the baseline (generic fitting) as a measure for comparison. (a) The comparison of the two versions of the pipeline, i.e. the online (incremental denoted as PD^2T_{INCREM}) and the offline, for two different generic fitting methods. (b) Contribution of the different pipeline steps (two different initialisation methods, i.e. a stronger one (SRDCF + CFSS) and a weaker one (MIL + CFSS)). The legend with P-S is an abbreviation of person-specific.

TABLE 2: (Preferably viewed in colour) Exemplar results with indicative fitting quality per experiment. The last column corresponds to errors considered as a failure, hence they are not visible in the CED plots and respectively not computed in the AUC. For the facial tracking the maximum error considered is 0.08, while for the animal tracking it is 12.

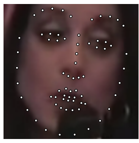
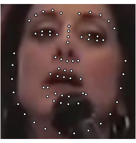
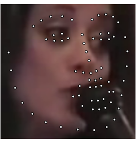
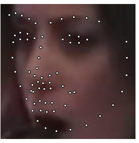
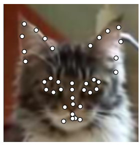
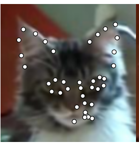
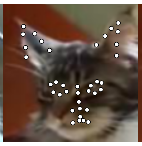
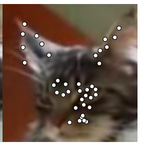
Exper.	*Error*			
	0.01	0.04	0.08	0.09
faces				
	2	7	12	13
cats				

TABLE 3: The hyperparameters used in the pipeline steps as utilised in the experiments, i.e. the facial landmark tracking, animal deformable tracking. Several hyperparameters are common in the two experiments due to less effort spent in optimising them, hence we reckon that improved results could be achieved with some tuning.

Category	*Hyper-parameters*	*Faces*	*Cats*
classifier	patch shape	(14, 14)	(12, 12)
	features	SIFT [37]	
	# pos. images	5000	2500
pictorial	patch shape	(9, 9)	(8, 8)
	features	sparse HOG [2]	
	# learning images ($\mathcal{K}_0$)	200	100
AAM	patch shape	[(11, 11), (17, 17)]	
	features	SIFT [38]	
	# learning images	30	

3.1 Technical details

PD^{2}T was implemented in Python with several utilities adopted from Menpo [35]. The software of [36] was used during the review process. In Tab. 3 some informative hyper-parameters are gathered, while additional details along with qualitative results[3] have been deferred to the supplementary material due to limited space; the rest technical details allowing the reproduction of the results are developed below.

The following preprocessing was performed for training the classifier: (a) A dataset of static images with the appropriate annotation, e.g. the trainset of the 300W challenge [39] for the facial tracking experiment, was utilised for extracting the positive training samples; perturbed versions of the annotations of those images along with selected images of Pascal dataset [40] were used for mining the negative samples. (b) For each positive sample a fixed size patch was extracted around each of the n landmark points, SIFT [37] were computed per patch. For each negative sample, a random perturbation of the ground truth points was performed to create an erroneous fitting prior to extracting the patches. (c) The linear SVM as a failure checking method was trained, with its hyper-parameters cross-validated in withheld validation clips (trainset of 300vW [9]).

We have noticed that that several challenging frames may contain outliers. In order to avoid having the covariance matrix affected by them, we used the trimmed statistical estimates [41]. Specifically, we have trimmed the top and bottom 3% of the values of the shape covariances.

In the person-specific localisation (Sec. 2.4), the number of generic shapes Q for the human face was $Q = 6$, while for cats

[3] In https://youtu.be/3QdWoTuJqgE we provide a video with the tracking results of several methods and different clips from our extensive comparisons.

$Q = 0$, i.e. no-generic shape was utilised for cats. These shapes are some shapes we experimentally noticed that appear rarely, e.g. closed-eyes or mouth wide open.

The computational cost of PD^{2}T is on-par with the most accurate techniques for generic fitting, which constitutes it an effective tool in case the quality of the landmarks is crucial. Specifically, following the report of [15] the mean time per frame (in seconds) is reported in Tab. 4. The generic results are copied from [15]. It should be noted that the bottleneck in our method is the generic fitting. The steps of learning the statistics for both the pictorial and the person-specific localisation are quite fast, in comparison to the respective learning for DPM and person-specific CFSS. Finally, we have created a GPU implementation of the incremental version of the algorithm which can track around 50 frames per second.

3.2 PD^{2}T Experimental Analysis

We assessed the performance of different steps and options of the proposed pipeline[4]. The two most crucial experiments developed below are a) the comparison of the online vs the offline version, b) the contribution of the different pipeline steps as developed above.

The performance of the two proposed versions, i.e. the incremental and the offline was scrutinised. In the incremental version, the first 100 frames were assumed existing and then the model was updated online every 30 frames. This validation was performed in four clips of the most challenging category of 300vW in two different generic results (one with very strong generic performance and one with mediocre). The two indicative methods were LCT [44] + CFSS [5] and MIL [43] + CFSS respectively. The CED plot in Fig. 4(a) demonstrates that the two version are equivalent with the offline version performing slightly better in the case of a weaker initial generic result. However, the strong benefit of the incremental version is the ability to execute in an online setting. In the rest of the manuscript, we focused on providing the optimal result that can be achieved currently, hence we have used the offline version.

We analysed the improvements from the different steps of PD^{2}T. The same setting as the comparison between the two versions was followed with MIL + CFSS (poor generic fitting) and SRDCF [45] + CFSS (strong generic fitting). In Fig. 4(b) the results of the output of steps 2 (person-specific detection + generic fitting) and 3 (person-specific localisation) are visualised. The performance gain in the two cases differs largely, since in the poor initial fitting the person-specific detection offers a huge boost, while in the SRDCF case with strong generic fitting, the gain originates mostly from the person-specific landmark localisation step. The experimental validation of our approach in two widely different initialisation conditions demonstrates the merit of learning both a person-specific detector and a person-specific localisation technique in order to ensure the optimal fitting.

3.3 Deformable facial sparse tracking

Tracking of facial parts has dominated the deformable object alignment literature, hence we investigated PD^{2}T with various generic fitting techniques. The dataset of 300vW [9] was utilised for sparse tracking. Additionally, a new dataset, FTOVM, covering cases that do not emerge in 300vW was annotated to validate our claim in different scenarios.

3.3.1 300 Videos in-the-Wild experiment

Using the 300 Videos in-the-Wild (300vW) dataset [9] the proposed pipeline in the task of face tracking was assessed. The dataset includes in total 114 lengthy videos (out of which 64 for testing) with each frame containing a single human face annotated with the sparse 68 mark-up. It is divided into the subsequent 3 categories:

- *Category 1*: Videos in well-lit environments without any occlusions.
- *Category 2*: Videos with unconstrained illumination conditions.
- *Category 3*: Video sequences captured in arbitrary conditions (including severe occlusions and extreme illuminations).

In total, 123404 frames are annotated in the testset of 300vW, consisting it the most extensive public dataset for evaluating the deformable facial tracking performance.

The recent comprehensive evaluation of [15] provided an excellent reference for generic methods that we utilised to verify that PD^{2}T works under a diverse set of conditions and initialisations, and provides a significant boost in state-of-the-art methods. It should be emphasised that all the hyper-parameters remained the same as in Tab. 3 for all the methods compared. To provide a fair comparison we report the same error metrics as in the original evaluation, i.e. the cumulative error distribution (CED) plots with normalised point-to-point error along with the Area Under the Curve (AUC) and the failure rate; in Fig. 2 some exemplar errors are plotted. Each error equals the mean Euclidean distance of the 68 points, normalised by the diagonal of the ground truth bounding box (further details for the errors exist on [15]). The combinations of methods (from [15]) were selected in order to provide as broad overview as possible, while as experimentally validated in [15] the landmark localisation method with the most accurate results for all combinations was CFSS [5]. Thus in the majority of the experiments we used CFSS, however we experimented with methods that use SDM [25]. Specifically, the methods chosen were the state-of-the-art detector of [1] (MTCNN); the correlation-filter based LCT [44]; the accurate discrimative DLSSVM of [46]; the discrimative SPT tracker [47]; the generative trackers of LSRST [48], DF [49] and RPT [50]; CMT [51] as a strong performing keypoint-based tracker; the widely used baselines of PF [52] and MIL [43]. Additionally, the baseline method that uses the tightest bounding box of the previous frame to initialise the current frame was included (denoted as PREV).

The quantitative results are provided in Table 5 and Fig. 5, while in Fig. 9 few qualitative fittings are visualised[3]. The outcomes demonstrate that PD^{2}T achieved the goal of improving the generic deformable tracking results in all methods and across all categories. Methods with weaker performance, e.g. DF + CFSS or SPT + CFSS gain a substantial performance boost that exceeds 50% of the initial AUC. This is particularly evident in category 3, which contains the most challenging videos. Nevertheless, for methods with stronger results, e.g. LRST + CFSS, the improvement is also noticeable in both the AUC, failure rate statistics and the CED plots.

As indicated in [15] (and Table 5) the first 2 categories are saturated with several methods performing equally well, thus our additional tests are restricted to the most challenging Category 3. The additional methods were: the state-of-the-art method of

[4] Considerable part of the self evaluation has been deferred to the supplementary material due to the restricted space.

DPM [2], [3]	MDNET [42]	MIL [43]	MTCNN [1]	CFSS [5]	PICTORIAL [27]	GN-DPM [34]	PD^{2}T (total)
2.087	3.101	0.075	3.204	1.207	1.121	0.093	2.531

TABLE 4: Computational cost of the top performing (generic and adaptive) methods. A generic deformable result has the computational cost of the detection + landmark localisation, i.e. MTCNN + CFSS. The cost is reported in seconds, i.e. the mean processing time for a single frame.

Method	Category 1			Category 2			Category 3		
	AUC	*Failure Rate (%)*	*Diff. (%)*	*AUC*	*Failure Rate (%)*	*Diff. (%)*	*AUC*	*Failure Rate (%)*	*Diff. (%)*
CMT + CFSS	0.746	2.681		0.755	1.886		0.592	16.676	
PD^{2}T(CMT + CFSS)	**0.762**	**4.611**	**2.145**	**0.787**	**1.326**	**4.238**	0.657	15.094	10.98
DF + CFSS	0.465	38.830		0.457	35.599		0.346	51.851	
PD^{2}T(DF + CFSS)	0.725	10.908	55.914	0.721	8.524	57.768	0.667	14.845	92.775
DLSSVM + CFSS	0.759	2.540		0.744	0.493		0.609	15.386	
PD^{2}T(DLSSVM + CFSS)	**0.779**	**3.591**	**2.635**	**0.786**	**1.232**	**5.645**	0.698	9.214	14.614
LCT + CFSS	0.704	10.551		0.767	0.636		0.641	12.930	
PD^{2}T(LCT + CFSS)	**0.767**	**5.566**	**8.949**	0.786	1.688	2.477	**0.707**	**10.503**	**10.296**
LRST + CFSS	0.702	10.890		0.756	1.621		0.646	13.588	
PD^{2}T(LRST + CFSS)	0.748	7.918	6.553	**0.790**	**1.050**	**4.497**	0.697	11.168	7.895
MIL + CFSS	0.681	11.497		0.707	4.229		0.378	45.985	
PD^{2}T(MIL + CFSS)	0.740	9.401	8.664	0.738	6.051	4.385	0.596	23.448	57.672
MTCNN + CFSS	0.729	8.578		0.720	8.527		**0.717**	**5.725**	
PD^{2}T(MTCNN + CFSS)	0.746	8.091	2.332	0.734	8.478	1.944	**0.741**	**5.439**	**3.347**
PF + CFSS	0.544	29.170		0.613	15.417		0.412	38.242	
PD^{2}T(PF + CFSS)	0.752	6.392	38.235	0.752	5.269	22.675	0.633	15.711	53.641
PREV + CFSS	0.543	27.472		0.616	19.883		0.198	73.052	
PD^{2}T(PREV + CFSS)	0.736	9.076	35.543	0.742	6.890	20.455	0.581	27.486	193.434
RPT + SDM	0.616	9.382		0.704	0.992		0.534	17.676	
PD^{2}T(RPT + SDM)	0.704	5.854	14.286	0.717	4.162	1.847	0.637	9.571	19.288
SPT + CFSS	0.266	64.091		0.215	67.629		0.156	76.758	
PD^{2}T(SPT + CFSS)	0.687	16.261	158.271	0.715	8.944	232.558	0.595	25.200	281.410
Colouring denotes the methods' performance ranking per category: 🟥 first 🟩 second 🟦 third									

TABLE 5: Results for Experiment of Section 3.3.1. For every method, we include the initial metrics and in the subsequent line the pipeline's metrics with that particular method as initialisation, denoted as PD^{2}T(method_name). The Area Under the Curve (AUC) and Failure Rate are reported. The last column per category, alleged Diff, denotes the improvement as a percentage of the AUC when applying PD^{2}T. The top performing methods (ranked by the AUC) of the table are highlighted.

[45] (SRDCF); the recent deep tracking system of [53] (SIAM-OXF); the two widely used methods of [31] (IVT), [54] (KCF). The additional results in Table 6 (Category 3) demonstrate that the performance gains are substantial. Notably, our method improves even the state-of-the-art systems, e.g. SRDCF + CFSS.

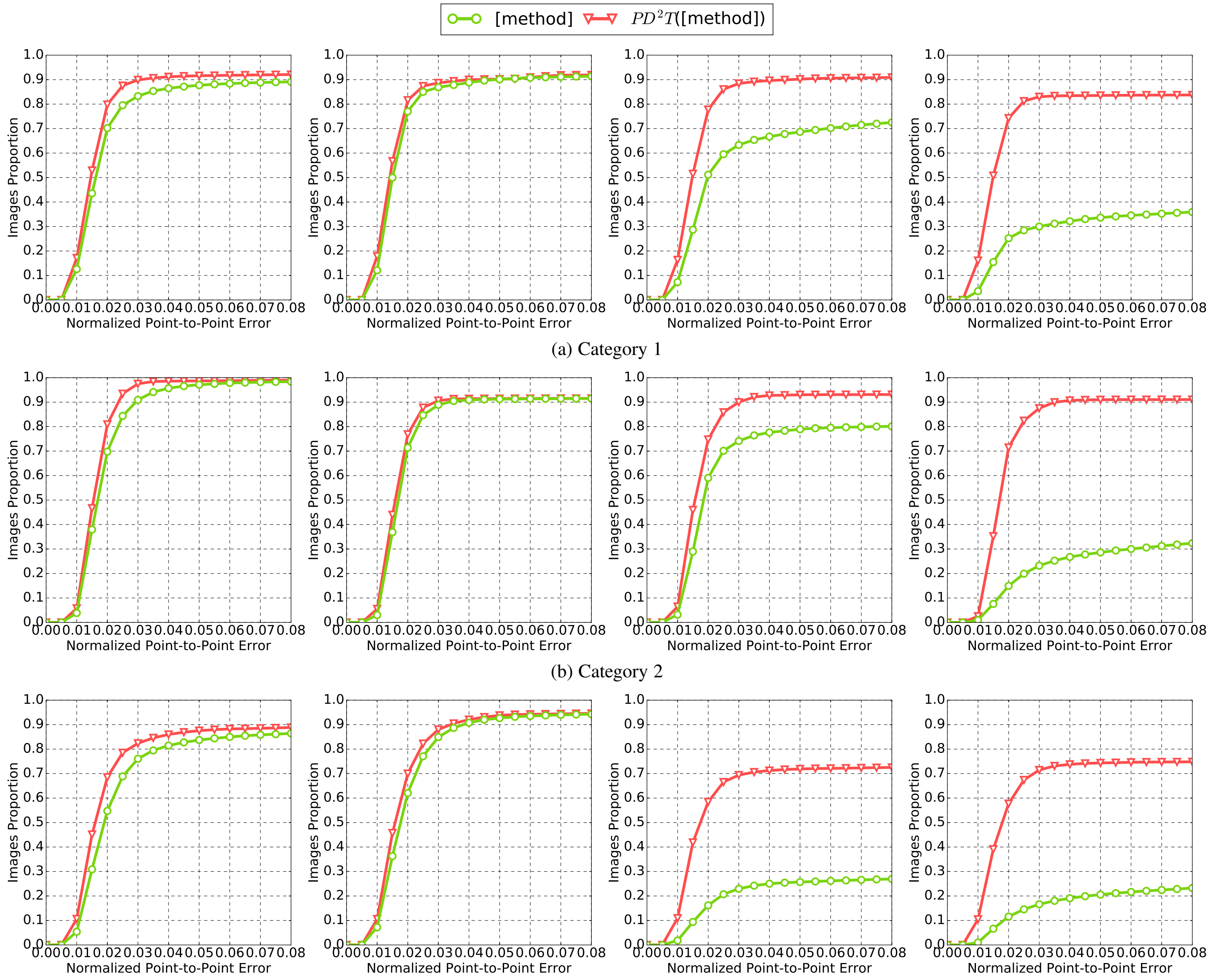
We additionally compared PD^{2}T against the state-of-the-art reported incremental tracker of [22]. We report the 3 top method outcomes of PD^{2}T against iCCR. The CED plot in Fig. 6 depicts the error metric as reported in the original paper, i.e. the error in the 49 points' markup normalised with the interocular distance. However, in constrast to the excluded frames of the original evaluation, we evaluated the errors in all the frames provided. All 3 outcomes from PD^{2}T outperform the results of iCCR; our method sets a new state-of-the-art for the most challenging category.

3.3.2 Facial Tracking with severe occlusion, out of view movement

Plenty of challenging cases are not included in 300vW. Namely in 300vW all the frames include exactly one face, while all the videos are of high resolution with the face close to the camera. The recent progress in tracking/detection systems dictates acquiring more challenging videos for experimentally assessing the merits of new systems. Thus, we have annotated semi-automatically 6 new clips that cover extreme cases that do not emerge in 300vW,

Method	*AUC*	*Failure Rate (%)*	*Difference (%)*
KCF + SDM	0.444	22.686	
PD^{2}T(KCF + SDM)	0.570	13.833	28.378
IVT + CFSS	0.423	42.244	
PD^{2}T(IVT + CFSS)	0.624	18.767	47.518
SIAM-OXF + SDM	0.567	13.775	
PD^{2}T(SIAM-OXF + SDM)	**0.634**	**13.877**	**11.816**
SRDCF + CFSS	**0.687**	**8.145**	
PD^{2}T(SRDCF + CFSS)	**0.749**	**4.955**	**9.025**
Colouring denotes the methods' ranking: 🟥 first 🟩 second 🟦 third			

TABLE 6: Results for Experiment with facial landmark tracking (Sec. 3.3.1). The connotation for the column legends and metrics are the same as in Fig. 5. The top 3 performing curves are highlighted.



(a) Category 1

(b) Category 2

(c) Category 3

Fig. 5: (Preferably viewed in colour) CED plots for Sec. 3.3.1 (deformable facial tracking). Depicted in (a) are the CED for Category 1, in (b) Category 2, in (c) Category 3. The common labelling method for all is that the green line denoted the initial result, while the red one the outcome of PD^{2}T. Two state-of-the art methods are depicted on the left, while two methods that benefit substantially from PD^{2}T are on the right two columns. Namely, from left to right the plots correspond to the methods: LRST + CFSS, MTCNN + CFSS, PREV + CFSS, SPT + CFSS. The complete set of plots is deferred to supplementary material due to limited space.

e.g. extreme occlusions, out of camera view movement, 360 degree rotation of the face. We name this new dataset FTOVM.

The annotation procedure for each clip was the following: The state-of-the-art MTCNN detector [1] was utilised, while the MDNET tracker was used to verify the existence of overlap between the tracked and the detected outcome. Subsequently, a generic GN-DPM [35], [55] was trained on 300W trainset and fit in all the frames. A human expert excluded the erroneous fittings; the GN-DPM model was augmented in an incremental manner and all the frames were re-fitted. A Kalman filter was used to smooth the fittings and a human expert kept only the frames with an accurate fitting.

The experimental setup remained the same as in 300vW; the generic deformable tracking method selected for the task was MDNET + CFSS. The comparison with PD^{2}T can be found in Fig. 7, while in the supplementary video qualitative results are visualised. The CED indicates that the strong method MDNET + CFSS drifts under severe occlusions or out of view movement, while our methods does not drift in such challenging cases.

3.4 Deformable animal tracking

Even though animals' faces include higher degrees of deformation and appearance variation, the subject has only recently emerged in static images, e.g. in [56], [57], however there are no established datasets for tracking yet. To illustrate the strengths of the proposed pipeline beyond human facial tracking, we annotated videos containing animal faces with an ad hoc shape. Six videos including cats' faces were semi-automatically annotated per frame with 38 facial markup; a generic GN-DPM [35], [55] was applied; sequentially a video-specific method [26] for refining the results was employed. The erroneous fittings were rejected by two human experts. Some clips contain more than one cats in order to verify the ability of the detector to discriminate between similar objects appearing in a clip.

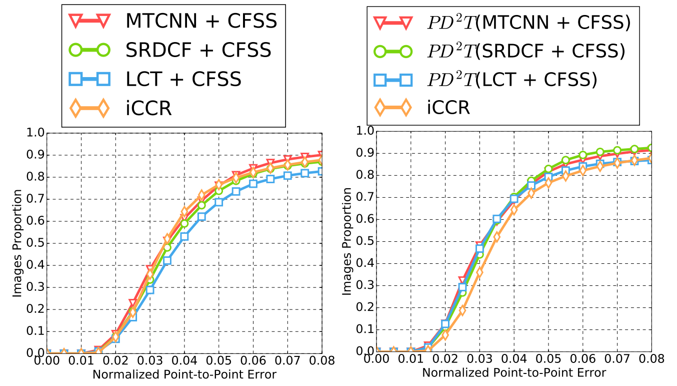


Fig. 6: (Preferably viewed in colour) CED plot for the comparison of iCCR against the best reported curves for PD^{2}T. In the left plot the generic methods are visualised against iCCR, while on the right the output of PD^{2}T with each respective initialisation. Apparently, iCCR is on-par with the generic MTCNN + CFSS; the two others are weaker, while on the right the PD^{2}T results in all three adaptive results surpassing iCCR.

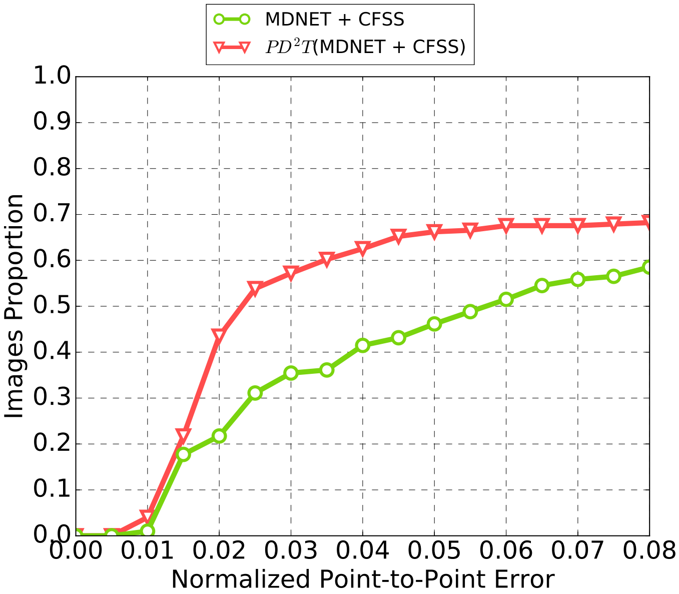


Fig. 7: (Preferably viewed in colour) CED plot for the newly introduced FTOVM in Sec. 3.3.2.

Three different experiments involving the two state-of-the-art model-free trackers of MDNET [42] and SRDCF [45] (first and second position respectively in the VOT 2015 challenge [16]) were performed. In the first two experiments, we obtained the bounding boxes from MDNET and varied the generic landmark localisation method, utilising SDM [25] and ERT [7], while on the last experiment we used SRDCF + SDM for the generic fitting. The specifics of SDM and ERT employed were the following: For SDM, a 4-level Gaussian pyramid with parametric shape was selected. SIFT [37] feature vectors of length 128 were extracted at the first 3 scales, using RootSIFT by [58], while the highest scale consisted of pixel intensities. For ERT the default parameters of [7] were utilised. Both SDM and ERT were trained on 2000 images of cat faces with the appropriate annotation, while the aforementioned parameters were validated in two withheld clips.

Due to the lack of established benchmark in the category, we resort to reporting the cumulative error plot that indicates in the x axis the point-to-point distance of the tracked point from the ground-truth (refer to Table 2 for indicated error values overlaid in the respective frame); additional metrics are deferred to supplementary material. The CED plots can be found in Fig. 8. The quantitative results in the generic method SRDCF + CFSS demonstrate that the improvement from an adaptive method are significant, while in the MDNET + CFSS, PD^{2}T outperforms the generic fitting with a large margin.

4 Conclusion

In this work, we introduced PD^{2}T, a pipeline for learning person-specific representations from a video clip. Given a set of generic fitting results as a (sparse) set of fiducial points, PD^{2}T aims at improving the generic fittings. From the internal evaluation of our method, it can be comprehended that the meticulous selection of the training samples for learning person-specific statistics results leads in very representative person-specific subspaces. We have conducted extensive experimentation with both human faces and animals' (cats) faces indicating that PD^{2}T improves the generic fitting results, while we have also demonstrated that the method works with a variety of different initial fittings' combinations. In the future we aim to explore different options for improving the results further, e.g. utilising features from recent deep networks, or learning a non-linear classifier for pruning the fittings. Additionally, we aim to explore the adaptation of our method for the case of missing data.

5 Acknowledgements

We would like to thank Epameinondas Antonakos and Patrick Snape for our fruitful conversations and their contributions in the preliminary version of this work. The work of Stefanos Zafeiriou has been partially funded by the FiDiPro program of Tekes (project number: 1849/31/2015), as well as the EPSRC project EP/N007743/1 (FACER2VM). The work of Grigorios Chrysos has been funded by a) the EPSRC project EP/L026813/1 Adaptive Facial Deformable Models for Tracking (ADAManT), as well as b) an Imperial College DTA.

References

[1] K. Zhang, Z. Zhang, Z. Li, and Y. Qiao, "Joint face detection and alignment using multi-task cascaded convolutional networks," *IEEE Signal Processing Letters*, vol. 23, no. 10, pp. 1499–1503, 2016, [Code: https://github.com/kpzhang93/MTCNN_face_detection_alignment, Status: Online; accessed 24-December-2016].

[2] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan, "Object detection with discriminatively trained part-based models," *IEEE Transactions on Pattern Analysis and Machine Intelligence (T-PAMI)*, vol. 32, no. 9, pp. 1627–1645, 2010.

[3] M. Mathias, R. Benenson, M. Pedersoli, and L. Van Gool, "Face detection without bells and whistles," in *Proceedings of European Conference on Computer Vision (ECCV)*. Springer, 2014, pp. 720–735.

[4] G. Trigeorgis, P. Snape, M. A. Nicolaou, E. Antonakos, and S. Zafeiriou, "Mnemonic descent method: A recurrent process applied for end-to-end face alignment," in *IEEE Proceedings of International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 4177–4187.

[5] S. Zhu, C. Li, C. C. Loy, and X. Tang, "Face alignment by coarse-to-fine shape searching," in *IEEE Proceedings of International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015, pp. 4998–5006, [Code: https://github.com/zhusz/CVPR15-CFSS, Status: Online; accessed 24-December-2016].

[6] J. Kossaifi, G. Tzimiropoulos, and M. Pantic, "Fast and exact bi-directional fitting of active appearance models," in *IEEE Proceedings of International Conference on Image Processing (ICIP)*. IEEE, 2015, pp. 1135–1139.

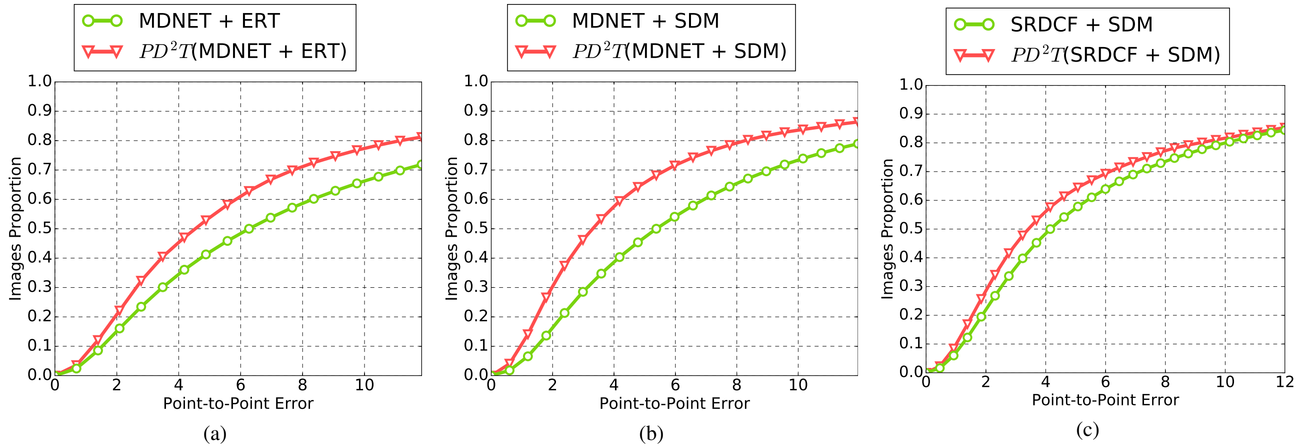
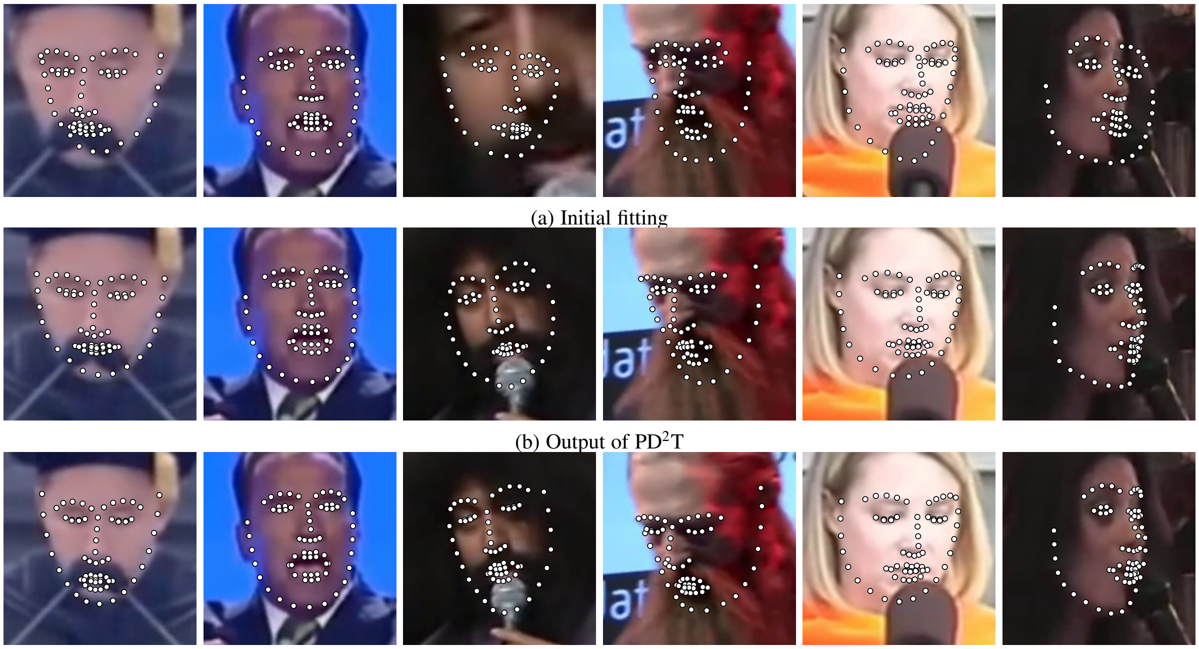


Fig. 8: (Preferably viewed in colour) CED plots for the animal tracking experiment (Sec. 3.4).



(a) Initial fitting

(b) Output of PD^2T

(c) Ground-truth annotation

Fig. 9: (Preferably viewed in colour) Indicative tracked frames from the experiment with the deformable facial tracking (Sec 3.3.1). The first row represents the generic fitting outcome (used as initialisation for PD^2T), the second row the outcome of PD^2T, and the third the ground-truth annotation. From left to right the initialisation methods were: IVT + CFSS, KCF + SDM, SIAM-OXF + SDM, SRDCF + CFSS, DLSSVM + CFSS, LRST + CFSS. Evidently, PD^2T results in a substantial improvement in the final localisation.

[7] V. Kazemi and J. Sullivan, "One millisecond face alignment with an ensemble of regression trees," in *IEEE Proceedings of International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014, pp. 1867–1874.

[8] R. A. Güler, G. Trigeorgis, E. Antonakos, P. Snape, S. Zafeiriou, and I. Kokkinos, "Densereg: Fully convolutional dense shape regression in-the-wild," 2017, pp. 6799–6808.

[9] J. Shen, S. Zafeiriou, G. Chrysos, J. Kossaifi, G. Tzimiropoulos, and M. Pantic, "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)*, December 2015, [Data: http://ibug.doc.ic.ac.uk/resources/300-VW/, Status: Online; accessed 9-November-2016].

[10] G. G. Chrysos and S. Zafeiriou, "Deep face deblurring," *IEEE Proceedings of International Conference on Computer Vision and Pattern Recognition Workshops (CVPR'W)*, 2017.

[11] X. Xiong and F. De la Torre, "Global supervised descent method," in *IEEE Proceedings of International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015, pp. 2664–2673.

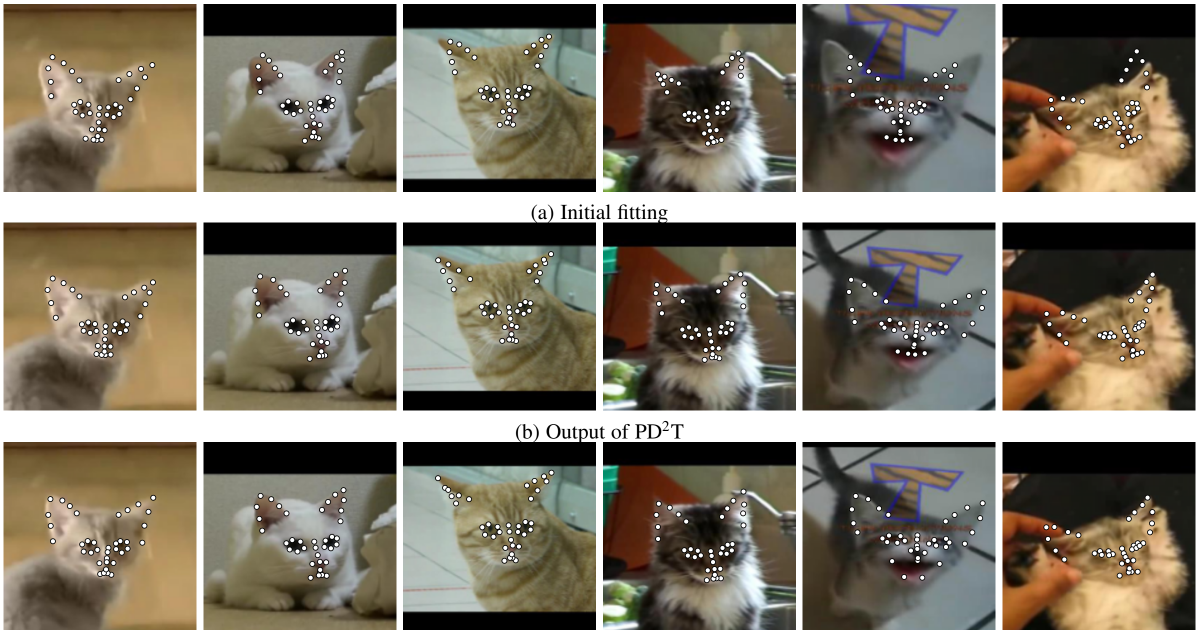
[12] A. Lanitis, C. J. Taylor, and T. F. Cootes, "A unified approach to coding and interpreting face images," in *IEEE Proceedings of International Conference on Computer Vision and Pattern Recognition (CVPR)*, 1995, pp. 368–373.

[13] S. Koelstra, M. Pantic, and I. Y. Patras, "A dynamic texture-based approach to recognition of facial actions and their temporal models," *IEEE Transactions on Pattern Analysis and Machine Intelligence (T-PAMI)*, vol. 32, no. 11, pp. 1940–1954, 2010.

[14] I. Essa, S. Basu, T. Darrell, and A. Pentland, "Modeling, tracking and interactive animation of faces and heads using input from video," in *Proceedings of Computer Animation*, 1996, pp. 68–79.

[15] G. G. Chrysos, E. Antonakos, P. Snape, A. Asthana, and S. Zafeiriou, "A comprehensive performance evaluation of deformable face tracking "in-the-wild"," *International Journal of Computer Vision (IJCV)*, 2017, [DOI: 10.1007/s11263-017-0999-5].

[16] M. Kristan, J. Matas, A. Leonardis, M. Felsberg, L. Čehovin, G. Fer-

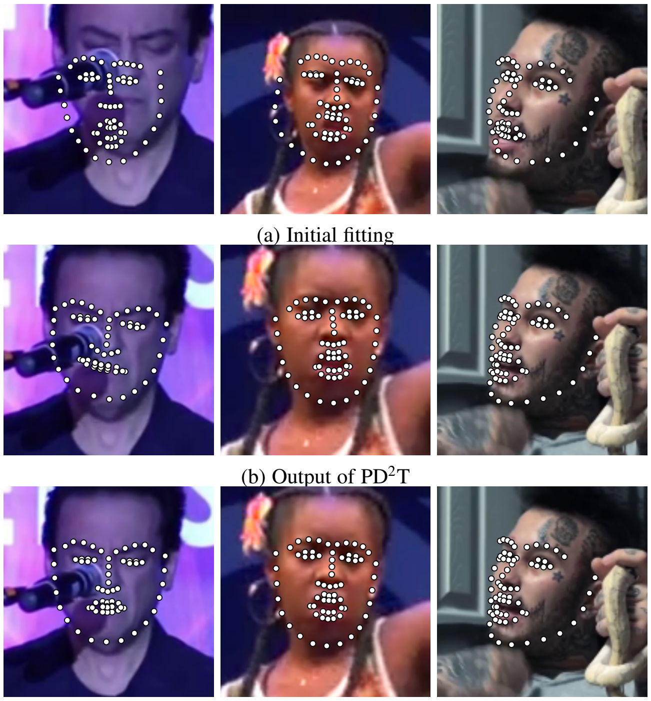


(a) Initial fitting

(b) Output of PD^{2}T

(c) Ground-truth annotation

Fig. 10: (Preferably viewed in colour) Indicative frames from the cats' facial tracking experiment of Sec 3.4).



(a) Initial fitting

(b) Output of PD^{2}T

(c) Ground-truth annotation

Fig. 11: (Preferably viewed in colour) Indicative tracked frames from the experiment with the deformable facial tracking with the newly introduced FTOVM (Sec 3.3.2). The first row represents the generic fitting (MDNET + CFSS), the second row the outcome of PD^{2}T; the third the ground-truth annotation.

nandez, T. Vojir, G. Häger, G. Nebehay *et al.*, "The visual object tracking vot2015 challenge results," in *IEEE Proceedings of International Conference on Computer Vision Workshops (ICCV'W)*, Dec 2015.

[17] Y. Wu, J. Lim, and M.-H. Yang, "Object tracking benchmark," *IEEE Transactions on Pattern Analysis and Machine Intelligence (T-PAMI)*, vol. 37, no. 9, pp. 1834–1848, 2015.

[18] Z. Chen, Z. Hong, and D. Tao, "An experimental survey on correlation filter-based tracking," *arXiv preprint arXiv:1509.05520*, 2015.

[19] Y. Peng, A. Ganesh, J. Wright, W. Xu, and Y. Ma, "Rasl: Robust alignment by sparse and low-rank decomposition for linearly correlated images," *IEEE Transactions on Pattern Analysis and Machine Intelligence (T-PAMI)*, vol. 34, no. 11, pp. 2233–2246, 2012.

[20] X. Cheng, S. Sridharan, J. Saraghi, and S. Lucey, "Anchored deformable face ensemble alignment," in *European Conference on Computer Vision*. Springer, 2012, pp. 133–142.

[21] C. Sagonas, Y. Panagakis, S. Zafeiriou, and M. Pantic, "Raps: Robust and efficient automatic construction of person-specific deformable models," in *IEEE Proceedings of International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014, pp. 1789–1796.

[22] E. Sánchez-Lozano, B. Martinez, G. Tzimiropoulos, and M. Valstar, "Cascaded continuous regression for real-time incremental face tracking," in *European Conference on Computer Vision*. Springer, 2016, pp. 645–661.

[23] X. Xiong and F. De la Torre, "Supervised descent method for solving nonlinear least squares problems in computer vision," *arXiv preprint arXiv:1405.0601*, 2014.

[24] A. Asthana, S. Zafeiriou, S. Cheng, and M. Pantic, "Incremental face alignment in the wild," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 1859–1866.

[25] X. Xiong and F. De la Torre, "Supervised descent method and its applications to face alignment," in *IEEE Proceedings of International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2013, pp. 532–539.

[26] G. G. Chrysos, E. Antonakos, S. Zafeiriou, and P. Snape, "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)*, 2015, pp. 1–9.

[27] P. F. Felzenszwalb and D. P. Huttenlocher, "Pictorial structures for object recognition," *International Journal of Computer Vision (IJCV)*, vol. 61, no. 1, pp. 55–79, 2005.

[28] P. Felzenszwalb and D. Huttenlocher, "Distance transforms of sampled functions," Cornell University, pp. 1963-2004, Tech. Rep., 2004.

[29] X. Zhu and D. Ramanan, "Face detection, pose estimation, and landmark localization in the wild," in *IEEE Proceedings of International Confer-*

ence on Computer Vision and Pattern Recognition (CVPR). IEEE, 2012, pp. 2879–2886, [Code: https://www.ics.uci.edu/~xzhu/face, Status: Online; accessed 2-June-2016].

[30] C. Cortes and V. Vapnik, "Support-vector networks," *Machine learning*, vol. 20, no. 3, pp. 273–297, 1995.

[31] D. A. Ross, J. Lim, R.-S. Lin, and M.-H. Yang, "Incremental learning for robust visual tracking," *International Journal of Computer Vision (IJCV)*, vol. 77, no. 1-3, pp. 125–141, 2008, [Code: http://www.cs.toronto.edu/~dross/ivt/].

[32] T. F. Cootes, C. J. Taylor, D. H. Cooper, and J. Graham, "Active shape models-their training and application," *Computer Vision and Image Understanding*, vol. 61, no. 1, pp. 38–59, 1995.

[33] J. Alabort-i Medina and S. Zafeiriou, "Bayesian active appearance models," in *IEEE Proceedings of International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014, pp. 3438–3445.

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

[35] J. Alabort-i-Medina, E. Antonakos, J. Booth, P. Snape, and S. Zafeiriou, "Menpo: A comprehensive platform for parametric image alignment and visual deformable models," in *Proceedings of ACM International Conference on Multimedia (ACM'MM)*. ACM, 2014, pp. 679–682, [Code: http://www.menpo.org/, Status: Online; accessed 9-November-2016].

[36] V. Koukis, C. Venetsanopoulos, and N. Koziris, "˜ okeanos: Building a cloud, cluster by cluster," *IEEE internet computing*, vol. 17, no. 3, pp. 67–71, 2013.

[37] D. G. Lowe, "Object recognition from local scale-invariant features," in *IEEE Proceedings of International Conference on Computer Vision (ICCV)*, 1999, pp. 1150–1157.

[38] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *IEEE Proceedings of International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2005, pp. 886–893.

[39] C. Sagonas, E. Antonakos, G. Tzimiropoulos, S. Zafeiriou, and M. Pantic, "300 faces in-the-wild challenge: Database and results," in *Image and Vision Computing*, 2015, [Data: http://ibug.doc.ic.ac.uk/resources/300-W/, Status: Online; accessed 15-January-2017].

[40] M. Everingham, L. Van Gool, C. K. Williams, J. Winn, and A. Zisserman, "The pascal visual object classes (voc) challenge," *International Journal of Computer Vision (IJCV)*, vol. 88, no. 2, pp. 303–338, 2010, [Data: http://host.robots.ox.ac.uk/pascal/VOC/voc2007/index.html, Status: Online; accessed 15-January-2017].

[41] R. Maronna, R. D. Martin, and V. Yohai, *Robust statistics*. John Wiley & Sons, Chichester. ISBN, 2006.

[42] H. Nam and B. Han, "Learning multi-domain convolutional neural networks for visual tracking," in *IEEE Proceedings of International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, [Code: https://github.com/HyeonseobNam/MDNet, Status: Online; accessed 9-November-2016].

[43] B. Babenko, M. H. Yang, and S. Belongie, "Robust object tracking with online multiple instance learning," *IEEE Transactions on Pattern Analysis and Machine Intelligence (T-PAMI)*, vol. 33, no. 8, pp. 1619–1632, August 2011.

[44] C. Ma, X. Yang, C. Zhang, and M.-H. Yang, "Long-term correlation tracking," in *IEEE Proceedings of International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015, pp. 5388–5396, [Code: https://github.com/chaoma99/lct-tracker, Status: Online; accessed 9-November-2016].

[45] M. Danelljan, G. Häger, F. Shahbaz Khan, and M. Felsberg, "Learning spatially regularized correlation filters for visual tracking," in *IEEE Proceedings of International Conference on Computer Vision (ICCV)*, 2015, pp. 4310–4318, [Code: https://www.cvl.isy.liu.se/en/research/objrec/visualtracking/regvistrack/, Status: Online; accessed 9-November-2016].

[46] J. Ning, J. Yang, S. Jiang, L. Zhang, and M.-H. Yang, "Object tracking via dual linear structured svm and explicit feature map," in *IEEE Proceedings of International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, [Code: www4.comp.polyu.edu.hk/~cslzhang/code/DLSSVM_CVPR.zip, Status: Online; accessed 18-August-2016].

[47] F. Yang, H. Lu, and M.-H. Yang, "Robust superpixel tracking," *IEEE Transactions in Image Processing (TIP)*, vol. 23, no. 4, pp. 1639–1651, 2014, [Code: http://www.umiacs.umd.edu/~fyang/spt.html, Status: Online; accessed 18-August-2016].

[48] T. Zhang, S. Liu, N. Ahuja, M.-H. Yang, and B. Ghanem, "Robust visual tracking via consistent low-rank sparse learning," *International Journal of Computer Vision (IJCV)*, vol. 111, no. 2, pp. 171–190, 2014, [Code: https://goo.gl/WiJb4z, Status: Online; accessed 2-June-2016].

[49] L. Sevilla-Lara and E. Learned-Miller, "Distribution fields for tracking," in *IEEE Proceedings of International Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2012, pp. 1910–1917, [Code: http://people.cs.umass.edu/~lsevilla/trackingDF.html, Status: Online; accessed 2-June-2016].

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

[51] G. Nebehay and R. Pflugfelder, "Clustering of Static-Adaptive correspondences for deformable object tracking," in *IEEE Proceedings of International Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2015, [Code: https://github.com/gnebehay/CppMT, Status: Online; accessed 2-June-2016].

[52] M. Isard and A. Blake, "Contour tracking by stochastic propagation of conditional density," in *Proceedings of European Conference on Computer Vision (ECCV)*, 1996, pp. 343–356, [Code: https://github.com/gnebehay/SIR-PF, Status: Online; accessed 23-December-2016].

[53] L. Bertinetto, J. Valmadre, J. F. Henriques, A. Vedaldi, and P. Torr, "Fully-convolutional siamese networks for object tracking," *arXiv preprint arXiv:1606.09549*, 2016, [Code:https://github.com/bertinetto/siamese-fc, Status: Online; accessed 22-December-2016].

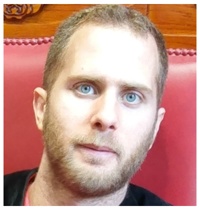
[54] J. F. Henriques, R. Caseiro, P. Martins, and J. Batista, "High-speed tracking with kernelized correlation filters," *IEEE Transactions on Pattern Analysis and Machine Intelligence (T-PAMI)*, vol. 37, no. 3, pp. 583–596, 2015, [Code: https://github.com/joaofaro/KCFcpp, Status: Online; accessed 2-June-2016].

[55] G. Tzimiropoulos and M. Pantic, "Gauss-newton deformable part models for face alignment in-the-wild," in *IEEE Proceedings of International Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2014.

[56] C. Sagonas, Y. Panagakis, S. Zafeiriou, and M. Pantic, "Robust statistical frontalization of human and animal faces," *International Journal of Computer Vision (IJCV)*, vol. 122, no. 2, pp. 270–291, 2017.

[57] A. Bulat and G. Tzimiropoulos, "Convolutional aggregation of local evidence for large pose face alignment," in *Proceedings of British Machine Vision Conference (BMVC)*, 2016.

[58] R. Arandjelović and A. Zisserman, "Three things everyone should know to improve object retrieval," in *IEEE Proceedings of International Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2012, pp. 2911–2918.



Grigorios G. Chrysos is a PhD student in iBUG group working with Stefanos Zafeiriou. He in his third year towards a PhD degree, while previously, he graduated from National Technical University of Athens (2014). He has since been working in the field of Computer Vision and statistical Machine Learning.

Stefanos Zafeiriou (M09) is currently a Reader in Machine Learning and Computer Vision with the Department of Computing, Imperial College London, London, U.K, and a Distinguishing Research Fellow with University of Oulu under Finish Distinguishing Professor Programme. He was a recipient of the Prestigious Junior Research Fellowships from Imperial College London in 2011 to start his own independent research group. He was the recipient of the Presidents Medal for Excellence in Research Supervision for 2016. He currently serves as an Associate Editor of the IEEE Transactions on Affective Computing and Computer Vision and Image Understanding journal. In the past he held editorship positions in IEEE Transactions on Cybernetics the Image and Vision Computing Journal. He has been a Guest Editor of over six journal special issues and co-organised over 13 workshops/special sessions on specialised computer vision topics in top venues, such as CVPR/FG/ICCV/ECCV (including four very successfully challenges run in ICCV13, ICCV15, CVPR17 and ICCV'17 on facial landmark localisation/tracking). He has co-authored over 55 journal papers mainly on novel statistical machine learning methodologies applied to computer vision problems, such as 2-D/3-D face analysis, deformable object fitting and tracking, shape from shading, and human behaviour analysis, published in the most prestigious journals in his field of research, such as the IEEE T-PAMI, the International Journal of Computer Vision, the IEEE T-IP, the IEEE T-NNLS, the IEEE T-VCG, and the IEEE T-IFS, and many papers in top conferences, such as CVPR, ICCV, ECCV, ICML. His students are frequent recipients of very prestigious and highly competitive fellowships, such as the Google Fellowship x2, the Intel Fellowship, and the Qualcomm Fellowship x3. He has more than 4500 citations to his work, h-index 36. He was the General Chair of BMVC 2017.